

# A multivariate approach for correcting reporting delays in infectious disease surveillance

Guilherme dos Santos



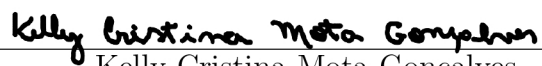
Universidade Federal do Rio de Janeiro  
Instituto de Matemática  
Departamento de Métodos Estatísticos

2023

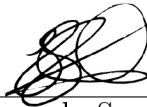
# A multivariate approach for correcting reporting delays in infectious disease surveillance

Guilherme dos Santos

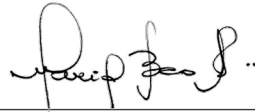
Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística. Aprovado por:



Kelly Cristina Mota Gonçalves  
Dr.Sc. - IM/UFRJ - Orientadora.



Leonardo Soares Bastos  
Dr.Sc. - PROCC/FIOCRUZ - Coorientador.



Mariane Branco Alves  
Dr.Sc. - IM/UFRJ.



Carlos Antonio Abanto-Valle  
Dr.Sc. - IM/UFRJ.



Paulo Inácio de Knecht López de Prado  
Dr.Sc. - IB/USP.

Rio de Janeiro, RJ - Brasil  
03 de abril de 2023

## CIP - Catalogação na Publicação

S237m Santos, Guilherme dos  
A multivariate approach for correcting reporting  
delays in infectious disease surveillance /  
Guilherme dos Santos. -- Rio de Janeiro, 2023.  
48 f.

Orientadora: Kelly Cristina Mota Gonçalves.  
Coorientador: Leonardo Soares Bastos.  
Dissertação (mestrado) - Universidade Federal do  
Rio de Janeiro, Instituto de Matemática, Programa  
de Pós-Graduação em Estatística, 2023.

1. Estatística. 2. Inferência Bayesiana. 3.  
Nowcasting. 4. Modelagem bayesiana hierárquica . 5.  
Arboviroses . I. Gonçalves, Kelly Cristina Mota,  
orient. II. Bastos, Leonardo Soares, coorient. III.  
Título.

# Abstract

Frequently, real-time tracking of epidemics is faced with a concerning issue, the reporting delays of cases and deaths. Delays might occur due to logistical problems, laboratory confirmation, and other reasons. Being able to correct the delay is essential to decision-making with the goal of containing an epidemic. In some cases, the epidemic might be associated with more than one disease, Dengue and Chikungunya are common examples of this phenomenon. We propose a multivariate model to correct reporting delays and accommodate the above-mentioned cases. The model is estimated using the Integrated Nested Laplace Approximation method with the aim of providing faster results.

*Keywords:* Nowcasting. Dengue. Chikungunya. INLA. Bayesian hierarchical model.

# Resumo

Frequentemente, o monitoramento em tempo real de epidemias enfrenta um problema preocupante: os atrasos na notificação de casos e mortes. Esses atrasos podem ocorrer devido a problemas logísticos, confirmação laboratorial e outros motivos. A habilidade de corrigir esses atrasos é essencial para a tomada de decisões com o objetivo de conter uma epidemia. Em alguns casos, a epidemia pode estar associada a mais de uma doença, como é o caso de dengue e chikungunya, por exemplo. Neste trabalho, propomos um modelo multivariado para corrigir os atrasos de notificação e acomodar os casos como dengue e chikungunya. O modelo é estimado usando o método INLA (Integrated Nested Laplace Approximation) com o objetivo de obter resultados mais rapidamente.

*Palavras-chave:* Nowcasting. Dengue. Chikungunya. INLA. Modelo hierárquico bayesiano.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Materials and Methods</b>	<b>5</b>
2.1	Data structure . . . . .	5
2.2	Model specification . . . . .	6
2.3	Parameter estimation and nowcasting . . . . .	7
2.4	Model evaluation criteria . . . . .	8
<b>3</b>	<b>Application</b>	<b>11</b>
3.1	Study with artificial dataset . . . . .	11
3.1.1	Data generation and application of proposed model . . . . .	11
3.1.2	Evaluating different models . . . . .	16
3.2	Dengue and chikungunya in Rio de Janeiro . . . . .	17
3.2.1	Sliding windows analysis . . . . .	17
3.2.2	Comparing model structures . . . . .	21
3.2.3	Prior sensitivity analysis . . . . .	25
3.3	Discussion . . . . .	28
<b>4</b>	<b>Concluding remarks</b>	<b>31</b>
<b>A</b>	<b>Supplementary material for the study with artificial dataset</b>	<b>35</b>
<b>B</b>	<b>Supplementary material for the sliding windows study</b>	<b>37</b>
<b>C</b>	<b>Comparison of univariate approach and proxy approach</b>	<b>38</b>
<b>D</b>	<b>Supplementary material for the prior sensitivity analysis</b>	<b>39</b>

## List of Figures

1.1	Number of actually reported (solid line) and without delay (dashed line) infections by dengue (blue) and chikungunya (red) in the state of Rio de Janeiro by week of symptom onset and disease from 2017 to 2019. . . . .	4
3.1	Simulated time series used in the study. . . . .	12
3.2	Real values, posterior median, and respective 95% credible intervals for the random effects in the proposed model. . . . .	14
3.3	Nowcasting estimates, data observed at the moment of predictions, and true observed data artificially generated in the simulation study. . . . .	15
3.4	Proportion of reported cases of dengue (Blue) and chikungunya (Red) by number of weeks of delay along with 1% of reported cases threshold (dashed line). . . . .	19
3.5	Nowcasting for each of the 25 windows of 70 weeks to which the model was fitted. . . . .	20
3.6	Number of tweets published in Rio de Janeiro mentioning the term “dengue” From 2017 to 2019. . . . .	21
3.7	Relative interval width, relative interval score, and MAPE for each of the 25 sliding windows colored by model. . . . .	23
3.8	Effects of time and delay for dengue (blue) and chikungunya (red). . . . .	24
3.9	Nowcasting using the Negative Binomial (panels (a) and (b)), and the Poisson (panels (c) and (d)) distributions considering full data until May 19th, 2019. . . . .	25
3.10	Posterior marginals for the hyperparameters according to prior of choice. . . . .	27
3.11	Nowcasting estimates and interval widths according to prior of choice. . . . .	28
3.12	Nowcasting estimates and interval widths for a moment of high number of cases according to prior of choice. . . . .	29
A.1	Standardized version of simulated data in Figure 3.1 . . . . .	35
A.2	Nowcasting estimates according to the Poisson model, data observed at the moment of predictions, and true observed data artificially generated in the simulation study. . . . .	36
A.3	Nowcasting estimates according to the independent model, data observed at the moment of predictions, and true observed data artificially generated in the simulation study. . . . .	36
B.1	Nowcasting of Dengue cases for each of the 25 windows of 70 weeks . . . . .	37
C.1	Effects estimates by model. . . . .	38
C.2	95% credible interval width according to model. . . . .	38

## List of Tables

2.1	Data structure in a reporting delay problem with the observed number of events (white cells), occurred-but-not-yet-reported number of events (light gray cells) and the future number of events that could be of interest to forecast (dark gray cells). . . . .	5
3.1	Hyperparameters used in the generation of the artificial dataset. . . . .	12
3.2	Point estimates and 95% credible intervals for fixed effects in the proposed model, with true values used in the generation of the artificial dataset. . . . .	13
3.3	Real values, point and interval estimates for the hyperparameters. . . . .	15
3.4	Performance results for models considered. . . . .	17
3.5	Cumulative percentage of reported cases for dengue and chikungunya by number of weeks of delay. . . . .	18
3.6	Window sizes considered, time taken to fit the model, and respective performance metrics, with optimum values in bold. . . . .	20
3.7	Models considered in the application and respective linear predictor structures. . . . .	22
3.8	Performance metrics for models used in the application. . . . .	22
3.9	Fixed effects estimates for M2 with full data. . . . .	24
3.10	Posterior estimates of hyperparameters for M2 with full data. . . . .	25
3.11	Priors considered for a sensitivity analysis . . . . .	26
D.1	Hyperparameter estimates according to prior of choice. . . . .	39



# Chapter 1

## Introduction

When dealing with infectious diseases, surveillance systems play a pivotal role in the assessment of the severity of an epidemic and the management of associated risks. The purpose of such systems is to provide accurate and timely information on the situation of a given disease to be monitored. One crucial feature of a good surveillance system is to reflect the current situation of the phenomenon of interest, yielding information as close to the present reality as possible. Ideally, such information is used immediately to aid in the decision-making process to detect and contain an outbreak, for instance.

Some examples of surveillance systems in Brazil that help public authorities to make informed decisions to disease control are the InfoDengue<sup>1</sup> and InfoGripe<sup>2</sup> systems. The InfoDengue system aggregates information and produces reports on dengue, zika and chikungunya. These diseases are arboviruses transmitted by the same mosquitoes, mainly *Aedes aegypti*. In Brazil, suspected cases of these diseases must be reported into a National Disease Notification System (SINAN). InfoGripe is an initiative to monitor severe acute respiratory infection (SARI) cases reported into the Influenza epidemiological surveillance system (SIVEP-Gripe) and was widely used during the COVID-19 pandemic (Bastos et al., 2020).

Although timeliness is of essential importance in disease surveillance, most systems deal with the pressing problem of reporting delays. That is, an event such as the onset of symptoms or death by a disease can often take several days or weeks to be recorded in the system. This can severely restrain the possibility of taking decisions based on the information currently available since the actual situation might be significantly different from the one reported in the system. Therefore, addressing this problem is essential to real-time tracking of infectious diseases and decision-making.

Reporting delays might occur due to logistical difficulties, laboratory confirmation, staffing constraints, and other reasons. From a statistical perspective, the problem of notification delay can be seen as a momentary censorship problem, where the observed data, such as the count of reported cases, will eventually be available. Note that there is an important distinction between observable and true data since there are cases that will never be reported (underreporting). The observable data are detected and will be eventually reported, while the true counts are the observed count plus the cases that will

---

<sup>1</sup><https://info.dengue.mat.br/>

<sup>2</sup><http://info.gripe.fiocruz.br/>

never be reported.

The task of estimating the actual observable data, and correcting for reporting delays, is referred to in the literature as nowcasting (a combination of “now” and “forecasting”). Statistical modeling approaches have been widely applied to nowcast diseases and increase the accuracy of surveillance systems around the world.

However, the notification delay problem is not restricted to epidemiological data, some approaches perform nowcasting to estimate the number of outstanding claims when there is a delay between the occurrence of the event of interest and the insurance claim (Renshaw and Verrall, 1998).

Usually, the delayed counts are thought of as a variable indexed by the time of occurrence, and the delay until reporting, denoted as  $n_{t,d}$  – the number of events that occurred at time  $t$  but were reported  $d$  units of time later. More details on the data structure, as well as notation, can be found in Chapter 2. The chain-ladder technique, proposed by Mack (1993) as a distribution-free method and used to estimate incurred-but-not-reported (IBNR) claims reserves in insurance settings, directly handles the distribution of  $n_{t,d}$ . In Renshaw and Verrall (1998) it was demonstrated that the underlying model of the chain-ladder technique is a generalized linear model with effects for time and delay. This technique has also been expanded to include parametric and nonparametric functional forms and potential covariates (England and Verrall, 2002; Barbosa and Struchiner, 2002).

More closely related to the correction of delays in disease surveillance, Bastos et al. (2019) propose a flexible hierarchical Bayesian model based on the chain-ladder technique that allows estimation of missing observable counts and prediction for future times. The model expands the chain-ladder technique to accommodate spatial effects and covariates, the inference is done using the Integrated Nested Laplace Approximation (INLA) method (Rue et al., 2009). Furthermore, applications to dengue and SARI in Brazil are performed.

With similar inference and application, Rotejanaprasert et al. (2020) present a model that takes into account a moving window for the nowcasting of dengue in Thailand while testing for window size and distribution of the counts. The inference is also done using INLA, and the Negative Binomial and Poisson distributions are considered.

The counts  $n_{t,d}$  can also be approached conditionally, by conditioning them on the total observed at time  $t$ . This conditional approach has been employed for nowcasting cases of *Escherichia coli* in Germany (Höhle and an der Heiden, 2014) and for detecting outbreaks in the presence of reporting delays (Salmon et al., 2015), for instance. In the latter case, the approach involves assuming a Negative Binomial distribution for the totals and a Multinomial distribution for the conditioned  $n_{t,d}$ . The model was implemented in INLA and is available in R via the `surveillance` package.

In more recent developments, Günther et al. (2021) use a two-step process to nowcast COVID-19 cases in Bavaria. Another proposed method is a Bayesian smoothing approach for nowcasting by McGough et al. (2020), which has been found to perform better in the presence of varying delays over time and has been tested on dengue data in Puerto Rico and influenza-like illness (ILI) in the United States. The approach is available through the R package `NobBS`.

In some cases, the epidemic might be associated with more than one disease. Dengue

and chikungunya are common examples of this phenomenon since these are associated with the same vector, the *Aedes aegypti* mosquito. Hence, it is reasonable to assume that epidemics of both diseases might occur concurrently. Figure 1.1 shows the number of cases of dengue and chikungunya in Rio de Janeiro from 2017 to 2019. It illustrates that there were moments with a simultaneously high number of cases of both diseases. The dashed line represents the data observed without delay, i.e. reported in the same week of the beginning of symptoms. Note that it hardly represents the real counts of cases.

In cases like the previously mentioned, it stands to reason that using a multivariate scheme to correct reporting delays could potentially bring benefits and improve the corrections. Dealing jointly with corrections might provide better handling of uncertainty. Stoner and Economou (2020) propose a multivariate approach based on a Generalized Dirichlet-Multinomial framework that allows for nowcasting and inclusion of underreporting in the final count of interest, but it is not suitable for approximation with INLA, for instance, demanding an MCMC approach.

In a setting of multivariate modeling of time series data. Berry and West (2020) present a multivariate dynamic generalized linear model for the forecasting of time series of non-negative counts with a decouple/recouple approach that allows for information sharing between series. Alves et al. (2022) tackle the estimation of dynamic generalized linear models within the k-parametric exponential family through a sequential approach based on information geometry.

In this work, we seek to propose a flexible multivariate Bayesian framework that allows the joint correction of delays for events that present some relationship, as in the case of infectious diseases such as dengue and chikungunya. We take inspiration from works that expand upon the chain-ladder technique, notably Bastos et al. (2019). We aim for a flexible and computationally cheap model without losing the ability to accommodate the complex nature of a problem that involves multivariate correction of delays. For the case counts, we assume a Negative Binomial distribution that has the Poisson distribution as a limit case. The model allows for common and marginal effects for time and amount of delay for each disease.

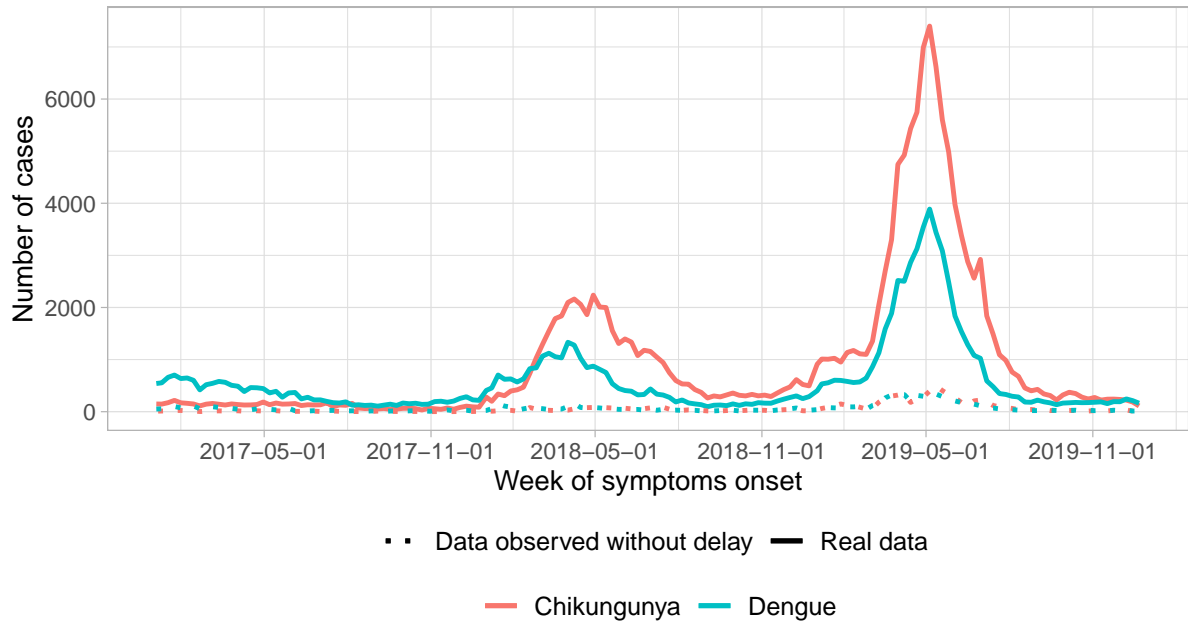


Figure 1.1: Number of actually reported (solid line) and without delay (dashed line) infections by dengue (blue) and chikungunya (red) in the state of Rio de Janeiro by week of symptom onset and disease from 2017 to 2019.

Since the posterior distributions of the parameter vector are not easily obtainable, it is necessary to use some method to sample from the resulting posterior distribution. Instead of using the usual Markov Chain Monte Carlo (MCMC) (Gamerman and Lopes, 2006) approach for bayesian inference, we perform approximate inference using the INLA method. This choice was made to obtain faster results and facilitate the future inclusion of the model in a surveillance system. The use of INLA in previous works such as Bastos et al. (2019) and Rotejanaprasert et al. (2020) also motivates the choice for approximate inference. INLA has been shown to provide very good approximations while reducing computation costs substantially (Rue et al., 2017) for Latent Gaussian Models (LGMs), which is the case of our proposed model, which is essentially a generalized mixed model.

The implementation and analysis shown in this work were performed in the R programming language v. 4.1.1 (R Core Team, 2021). The use of INLA in R is made possible by the R-INLA<sup>3</sup> package.

This document is organized as follows. In Chapter 2, we present the methodology, such as the data structure and notation, the model, and the inference procedure based on the posterior predictive distribution of the true observed count of events. In Chapter 3 we present an application of the model to simulated data and the data shown in Figure 1.1 from dengue and chikungunya arboviruses in Rio de Janeiro from 2017 to 2019 as well as a brief discussion. Concluding remarks are presented in Chapter 4.

---

<sup>3</sup><https://www.r-inla.org/>

# Chapter 2

## Materials and Methods

### 2.1 Data structure

The usual data structure in a reporting delay problem is shown in Table 2.1, where the rows are indexed by the time steps  $t = 1, 2, \dots, T+H$  and the columns by the amount of delay  $D = 0, 1, \dots, D$ . Here,  $T$  represents the current time and is usually measured in weeks in disease surveillance data, and  $D$  is the maximum acceptable delay, measured in the same unit as the time steps. The  $n_{t,d}$  cell represents the number of events that occurred at time  $t$  but were reported  $d$  units of time later.

$t \backslash d$	0	1	2	$\dots$	$D-2$	$D-1$	$D$	$N$
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,D-2}$	$n_{1,D-1}$	$n_{1,D}$	$N_1$
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,D-2}$	$n_{2,D-1}$	$n_{2,D}$	$N_2$
3	$n_{3,0}$	$n_{3,1}$	$n_{3,2}$	$\dots$	$n_{3,D-2}$	$n_{3,D-1}$	$n_{3,D}$	$N_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-D$	$n_{T-D,0}$	$n_{T-D,1}$	$n_{T-D,2}$	$\dots$	$n_{T-D,D-2}$	$n_{T-D,D-1}$	$n_{T-D,D}$	$N_{T-D}$
$T-D+1$	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$n_{T-D+1,2}$	$\dots$	$n_{T-D+1,D-2}$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$	$N_{T-D+1}$
$T-D+2$	$n_{T-D+2,0}$	$n_{T-D+2,1}$	$n_{T-D+2,2}$	$\dots$	$n_{T-D+2,D-2}$	$n_{T-D+2,D-1}$	$n_{T-D+2,D}$	$N_{T-D+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-2$	$n_{T-2,0}$	$n_{T-2,1}$	$n_{T-2,2}$	$\dots$	$n_{T-2,D-2}$	$n_{T-2,D-1}$	$n_{T-2,D}$	$N_{T-2}$
$T-1$	$n_{T-1,0}$	$n_{T-1,1}$	$n_{T-1,2}$	$\dots$	$n_{T-1,D-2}$	$n_{T-1,D-1}$	$n_{T-1,D}$	$N_{T-1}$
$T$	$n_{T,0}$	$n_{T,1}$	$n_{T,2}$	$\dots$	$n_{T,D-2}$	$n_{T,D-1}$	$n_{T,D}$	$N_T$
$T+1$	$n_{T+1,0}$	$n_{T+1,1}$	$n_{T+1,2}$	$\dots$	$n_{T+1,D-2}$	$n_{T+1,D-1}$	$n_{T+1,D}$	$N_{T+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T+H$	$n_{T+H,0}$	$n_{T+H,1}$	$n_{T+H,2}$	$\dots$	$n_{T+H,D-2}$	$n_{T+H,D-1}$	$n_{T+H,D}$	$N_{T+H}$

Table 2.1: Data structure in a reporting delay problem with the observed number of events (white cells), occurred-but-not-yet-reported number of events (light gray cells) and the future number of events that could be of interest to forecast (dark gray cells).

We adopt the notation employed by Höhle and an der Heiden (2014) and Bastos et al. (2019). At any given time point  $t$ , the number of events of interest (e.g., number of cases in a particular week) is the total  $N_t$ , given by the sum of the row  $t$  in the table. Note that

if  $t + d > T$ , the value of  $n_{t,d}$  has not yet been observed, and thus we cannot compute  $N_t$  without estimating such missing values. These unobserved  $n_{t,d}$  are represented by the light gray cells in Table 2.1, the so-called run-off triangle, and are the quantities of interest in a nowcasting problem, the occurred-but-not-yet-reported events. Additionally, if one is interested in forecasting  $H$  time steps ahead, the focus would be on the dark gray cells. In this work, we focus on nowcasting since the aim is to correct reporting delays in disease surveillance data.

When dealing with more than one disease our data can be seen as replicates of Table 2.1, one for each disease. Thus, we are interested in the totals  $N_{i,t}$ , and the cells of the table can be denoted as  $n_{i,t,d}$ , where  $i = 1, \dots, p$  indicates the  $i$ -th disease.

## 2.2 Model specification

Since we are dealing with more than one disease, we can now see our data as a replicate of Table 2.1 for each disease. Thus, let  $n_{i,t,d}$  be the count of the event of interest for  $i$ -th disease that occurred at time step  $t$  but were reported  $d$  units of times later, for  $t = 1, \dots, T$ ,  $d = 1, \dots, D$ , and  $i = 1, \dots, p$ . We assume that  $n_{i,t,d}$  follows a Negative Binomial distribution and the linear predictor combines joint and marginal effects, i.e.,

$$n_{i,t,d} \sim \text{NegBin}(\lambda_{i,t,d}, \phi_i), \quad \lambda_{i,t,d} > 0, \quad \phi_i > 0, \quad (2.1)$$

$$\log(\lambda_{i,t,d}) = \alpha_i + \beta_{i,t} + \gamma_{i,d} + \delta_t + \psi_d + \boldsymbol{\nu}_{i,t} \mathbf{x}_t, \quad (2.2)$$

for  $i = 1, \dots, p$ ;  $t = 1, \dots, T$ ; and  $d = 1, \dots, D$ .

We use the following parametrization for the Negative Binomial distribution:

$$\text{If } Y \sim \text{NegBin}(\lambda, \phi) \Rightarrow p(y|\lambda, \phi) = \binom{y + \phi - 1}{y} \left( \frac{\lambda}{\lambda + \phi} \right)^y \left( \frac{\phi}{\lambda + \phi} \right)^\phi, \quad y = 0, 1, 2, \dots \quad (2.3)$$

In this parametrization for the Negative Binomial distribution, the expected value and variance are given, respectively, by  $E(n_{i,t,d}) = \lambda_{i,t,d}$  and  $V(n_{i,t,d}) = \lambda_{i,t,d} (1 + \lambda_{i,t,d}/\phi_i)$ . Note that the Poisson distribution is obtained if we take the limit as  $\phi$  tends to infinity. Hence, the Negative Binomial distribution can be seen as a more flexible extension of the Poisson distribution that can accommodate overdispersion, with  $\phi$  accounting for the inverse of the overdispersion.

In Equation (2.2),  $\alpha_i$  is an intercept term accounting for the overall mean at log-scale, the effects  $\beta_{i,t}$  and  $\gamma_{i,d}$  capture the temporal evolution and the structure of the delay mechanism, respectively, for the  $i$ -th disease count. The terms  $\delta_t$  and  $\psi_d$  capture the common structures in time and delay, respectively, for both counts of interest. These effects are important to take the correlation between diseases into account. Moreover,  $\mathbf{x}_t$  is a vector of regressor variables and  $\boldsymbol{\nu}_{i,t}$  are the respective coefficients of these external variables varying over time and per disease. A particular case might be adopted when  $\boldsymbol{\nu}_i$  is static, where  $\mathbf{x}_t$  still varies in time. It is also possible to set a common value  $\boldsymbol{\nu}_i = \boldsymbol{\nu}$  for the coefficient if it seems suitable that the covariate in question influences both diseases equally. The random effect components have been constrained to sum up to zero to

ensure the identifiability of the fixed effects such as the intercept terms and the mean effect of the covariate.

In this specification, the effects of time and delay might evolve according to a variety of processes, such as random walks or autoregressive processes of a given order. In particular, a first order random walk for these effects is a possible choice and will be further used, that is,

$$\beta_{i,t} | (\beta_{i,t-1}, \sigma_\beta^2) \sim N(\beta_{i,t-1}, \sigma_\beta^2), \quad (2.4)$$

$$\gamma_{i,d} | (\gamma_{i,d-1}, \sigma_\gamma^2) \sim N(\gamma_{i,d-1}, \sigma_\gamma^2), \quad (2.5)$$

$$\delta_t | (\delta_{t-1}, \sigma_\delta^2) \sim N(\delta_{t-1}, \sigma_\delta^2), \quad (2.6)$$

$$\psi_d | (\psi_{d-1}, \sigma_\psi^2) \sim N(\psi_{d-1}, \sigma_\psi^2), \quad (2.7)$$

$$\nu_{i,t} | (\nu_{i,t-1}, \sigma_\nu^2) \sim N(\nu_{i,t-1}, \sigma_\nu^2), \quad (2.8)$$

for  $i = 1, \dots, p$ ;  $t = 2, \dots, T$ ; and  $d = 1, \dots, D$ .

Observe that the time effects evolve over time index  $t$  and the delay effects evolve over delay index  $d$ . Since the inference is done under a Bayesian approach, the model must be completed with the prior distributions. We can set, for instance, Gamma or Half-Normal priors for the precision hyperparameters  $\sigma^{-2}$  of each random walk. These priors are also suitable for  $\phi_i$ . It is worth noting that time series of infection counts might have longer temporal memory, which can be included in the model through higher-order random walks. Moreover, we consider prior independence for the effects.

## 2.3 Parameter estimation and nowcasting

We are interested in the posterior distribution of

$$\Theta = (\{\alpha_i\}, \{\beta_{i,t}\}, \{\gamma_{i,d}\}, \{\delta_t\}, \{\psi_d\}, \{\nu_{i,t}\}, \{\phi_i\}, \sigma_\beta^{-2}, \sigma_\gamma^{-2}, \sigma_\delta^{-2}, \sigma_\psi^{-2}, \sigma_\nu^{-2}).$$

which is given by

$$p(\Theta | \mathbf{n}) \propto p(\Theta) \prod_{i=1}^p \prod_{t=1}^T \prod_{\substack{d=0 \\ \{t+d < T\}}}^D p(n_{i,t,d} | \Theta), \quad (2.9)$$

where  $\mathbf{n} = \{n_{i,t,d}, t + d < T\}$  represents the observed data,  $p(n_{i,t,d} | \Theta)$  is the Negative Binomial probability function,  $p(\Theta)$  is the prior distribution given by the product of the marginal priors for the effects and hyperparameters.

In this case, the posterior is not analytically computable, which means there is a need for methods to obtain samples or an approximation from the posterior distribution of the parameters. To avoid the expensiveness of MCMC methods and facilitate the practical repeated use of the model (say, weekly in a surveillance system), we perform approximate inference through INLA.

INLA provides results significantly faster than MCMC approaches most of the time (Rue et al., 2017) and requires a less extensive examination of results such as convergence and autocorrelation analysis in MCMC. In short, INLA performs a sequence of Laplace

approximations and numerical methods for sparse matrices and is suitable for latent Gaussian models (which is the case for the proposed model). Hence, the model can be promptly implemented in R through R-INLA. The common procedure is to obtain the approximated posterior and sample from it with the `inla.sample.posterior()` function. After that step, we can sample  $n_{t,d}$  from the likelihood evaluated at each posterior sample.

Once we learn about the parameters through the posterior distribution, we can access the posterior predictive distribution of the occur-but-not-yet observed  $\{n_{t,d}; T < t + d < T + D\}$  to perform the nowcast. The posterior predictive is given by

$$p(n_{t,d} | \mathbf{n}) = \int_{\Theta} p(n_{t,d} | \Theta) p(\Theta | \mathbf{n}) d\Theta. \quad (2.10)$$

The integral in (2.10) does not have an analytical solution, however, an approximation is obtainable using a Monte Carlo approach. In practice, we take samples from the posterior  $p(\Theta | \mathbf{n})$  and then, for each parameter vector  $\theta$  sampled from  $p(\Theta | \mathbf{n})$ , we sample from the negative binomial likelihood  $p(n_{t,d} | \theta)$ .

Once the simulated values from the posterior predictive distribution are obtained, we can calculate the total number of notifications at time  $t$ ,  $N_t$ , by summing over the rows of Table 2.1. Some values will be known, and others will be taken from the posterior predictive samples.

In short, these are the steps taken to obtain the posterior predictive using the Monte Carlo approach after finding the approximation:

1. Sample  $(\{\alpha_i\}, \{\beta_{i,t}\}, \{\gamma_{i,d}\}, \{\delta_t\}, \{\psi_d\}, \{\nu_{i,t}\})$  from the joint posterior;
2. Sample  $n_{t,d}$  from the likelihood evaluated at the sampled parameters;
3. Compute  $N_t = \sum_{d=0}^D n_{t,d}$ .

Note that, in step 3, some  $n_{t,d}$  in the sum have already been observed, while others have been sampled in step 2.

In further applications in Chapter 3 we use vague gamma priors for the hyperparameters  $\{\{\phi_i\}, \sigma_\beta^{-2}, \sigma_\gamma^{-2}, \sigma_\delta^{-2}, \sigma_\psi^{-2}, \sigma_\nu^{-2}\}$ . All of these are set to follow a gamma distribution with parameters  $\alpha = 0.01$  and  $\beta = 0.01$ , leading to prior mean equals 1 and variance equals 100.

## 2.4 Model evaluation criteria

To assess the model performance, we considered a few metrics for both point and interval estimates. For point estimates, we chose to use the widely applied mean absolute percentage error (MAPE), mean squared error (MSE), and mean absolute error (MAE). In order to simplify the notation we will refer to the counts as  $y$ . Let  $\hat{y}_k$  and  $y_k$  be the



estimate and the actual value we aim to nowcast, respectively, the metrics mentioned above are given by the following expressions:

$$\begin{aligned}\text{MSE}(y_k, \hat{y}_k) &= \frac{1}{K} \sum_{k=1}^K (y_k - \hat{y}_k)^2, \\ \text{MAE}(y_k, \hat{y}_k) &= \frac{1}{K} \sum_{k=1}^K |y_k - \hat{y}_k|, \\ \text{MAPE}(y_k, \hat{y}_k) &= \frac{1}{K} \sum_{k=1}^K \left| \frac{y_k - \hat{y}_k}{y_k} \right| \times 100\%,\end{aligned}$$

where  $K$  denotes the number of steps predicted, or, in this case, nowcasted. Note that, for each of the metrics, the smaller the better.

For the interval estimates we aim for some metrics that value both width of the intervals and the ability to cover the real value. The interval score is one of the most common choices in this case. Let  $y$  be the count we aim to nowcast, and define  $l$  and  $u$ , respectively, as the lower and upper limits of an interval of credibility of level  $1 - \alpha$ . The interval score is given by the following expression:

$$\text{IS}_\alpha(y, l, u) = (u - l) + \frac{2}{\alpha} \times (l - y) \times \mathbb{1}(y < l) + \frac{2}{\alpha} \times (y - u) \times \mathbb{1}(y > u),$$

where  $\mathbb{1}$  is the indicator function. Note that the width of the interval is taken into account and there are also penalization terms for when the real value falls outside the interval. Smaller values of interval score point to better interval estimates. More information on the interval score and a visualization of the penalization can be found in Gneiting and Raftery (2007) and Bracher et al. (2021). We also measure the width and coverage of the intervals as a separate metric:

$$\text{Width}(l, u) = u - l.$$

We also define the coverage of interval estimates as

$$\text{Coverage}(y, l, u) = \frac{1}{K} \sum_{k=1}^K \mathbb{1}(l \leq y_k \leq u).$$

The previous metrics are useful and frequently used for univariate problems. In the case of a multivariate output, it can be cumbersome to examine the metrics for each series separately. As an attempt to create a unified metric rather than assessing individual time series separately, we chose to examine relative versions of interval score and interval width. These were obtained by normalizing both metrics with respect to the real value in the following manner:

$$\text{relIS}_\alpha(\mathbf{y}, \mathbf{u}, \mathbf{l}) = \sum_{i=1}^p \frac{\text{IS}_\alpha(y_i, l_i, u_i)}{|y_i|},$$

$$\text{relWidth}(\mathbf{y}, \mathbf{u}, \mathbf{l}) = \sum_{i=1}^p \frac{\text{Width}(y_i, l_i, u_i)}{|y_i|},$$

where  $i$  represents the  $i$ -th disease,  $y_i$  is the count we aim to nowcast,  $l_i$  and  $u_i$  are respectively the lower and upper limits of an interval of credibility of level  $1 - \alpha$ . Thus,  $\mathbf{y} = (y_1, \dots, y_p)$ ,  $\mathbf{l} = (l_1, \dots, l_p)$  and  $\mathbf{u} = (u_1, \dots, u_p)$ . Note that we are summing over the dimensions of a multivariate observation at time  $t$ . That is, in a disease reporting delay correction problem, the sum happens across the counts  $N_{i,t}$  for diseases observed at the same point in time.

We also evaluate two multivariate scoring rules for probabilistic forecasts that consider finite samples from a distribution of interest  $F$ , which in our case is the predictive posterior. The first metric is the energy score (ES; Gneiting and Raftery 2007), a multivariate generalization of the continuous ranked probability score (CRPS). The main idea behind the energy score is to measure the mean distance between the samples from the distribution of interest and the actual value and compare them with the mean distance inside the samples themselves. Smaller differences between the mean distances indicate better predictions.

The second metric is the variogram score (VS; Scheuerer and Hamill 2015), a proper scoring rule based on the geostatistical concept of variograms. Essentially, the variogram score measures the differences between the variograms of order  $\beta$  for the real data and the samples over all pairs of components of the multivariate vector. Smaller values indicate better performance. Let  $m$  be the number of samples from the distribution of interest  $F$ , the expressions for the energy score and variogram score are as follows:

$$\text{ES}(F, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \|\mathbf{X}_i - \mathbf{y}\| - \frac{1}{2m^2} \sum_{i=1}^m \sum_{j=1}^m \|\mathbf{X}_i - \mathbf{X}_j\|,$$

$$\text{VS}^\beta(F, \mathbf{y}) = \sum_{i=1}^p \sum_{j=1}^p w_{i,j} \left( |y^{(i)} - y^{(j)}|^\beta - \frac{1}{m} \sum_{k=1}^m |X_k^{(i)} - X_k^{(j)}|^\beta \right)^2,$$

where  $\mathbf{X}$  represents the vector samples taken from the posterior predictive and  $\mathbf{y}$  is the real observed vector. These and other metrics for probabilistic forecasts, although mostly univariate, are implemented and can be readily used through the R package `scoringRules` (Jordan et al., 2019). In further applications, we use the variogram score with weights  $w_{i,j} = 1$  and  $\beta = 0.5$ , which presented better discrimination against miscalibration in Scheuerer and Hamill (2015).

It is worth mentioning that variogram and energy scores are originally defined using the expectation with respect to the multivariate distribution  $F$  instead of the sample means. However, for finite samples, it is common to use the latter as it is shown here.

# Chapter 3

## Application

### 3.1 Study with artificial dataset

In this section, we evaluate the proposed model using artificially generated data. Subsection 3.1.1 covers the processes underlying data generation, the application of the proposed model, and its ability to recover the real values for the effects and hyperparameters. In subsection 3.1.2, we compare the nowcasting results of the proposed model to those of two other models.

#### 3.1.1 Data generation and application of proposed model

In order to validate the proposed model and its capacity to estimate delay and time effects, we propose an application with a simulated dataset that considers two time series  $N_1$  and  $N_2$ , representing two disease counts that have some similarity in their evolution. We generate the effects according to random walks of order 1, as explicit in Equations (2.4) to (2.8), except for the common delay effect  $\psi_d$ , which was modified to imitate the behavior of the real data. The evolution of the common delay effect is governed by Equation (3.1), that is, the effect for a delay of  $d$  units of time is set to follow a truncated normal distribution centered and truncated at the previous value, ensuring a decreasing effect over the amount of delay, this is:

$$\psi_d \sim \mathcal{TN}_{(-\infty, \psi_{d-1})}(\psi_{d-1}, \sigma_\psi^2). \quad (3.1)$$

In this study we use one covariate  $x$ , taking the form of a scalar with the behavior dictated by a random walk of order 1. The coefficient  $\nu_t$  is also considered to follow a random walk of order 1, and in this study, we consider a common coefficient for both time series. The true values of the inverse overdispersion hyperparameters  $\phi_i$ ,  $i = 1, 2$  and the precision hyperparameters for Equations (2.4) to (2.8) were fixed according to the values presented in Table 3.1.

Hyperparameter	Real value
$\phi_1$	5
$\phi_2$	5
$\sigma_\beta^{-2}$	50
$\sigma_\gamma^{-2}$	50
$\sigma_\delta^{-2}$	5
$\sigma_\psi^{-2}$	5
$\sigma_\nu^{-2}$	1/2

Table 3.1: Hyperparameters used in the generation of the artificial dataset.

The study considered a window of  $T = 200$  time steps and a maximum possible delay of  $D = 15$  time steps. Both generated time series are depicted in Figure 3.1. The scale of the y-axis is due to the difficulty of generating more well-behaved values once we are working in the logarithmic scale when generating the effects of time, delay, and covariate and computing the linear predictor. Nonetheless, some correlation between moments of peak can still be observed. Notably, the selection of a random walk of order one to the time effect led to a less smooth curve than in real applications, which may theoretically impose greater difficulty on the nowcasting process. In summary, the generated dataset presents a behavior comparable to or more challenging than the real dataset, which motivates this work. To facilitate a clearer analysis of the relation between both time series, Appendix A is included, containing a standardized representation of the generated data.

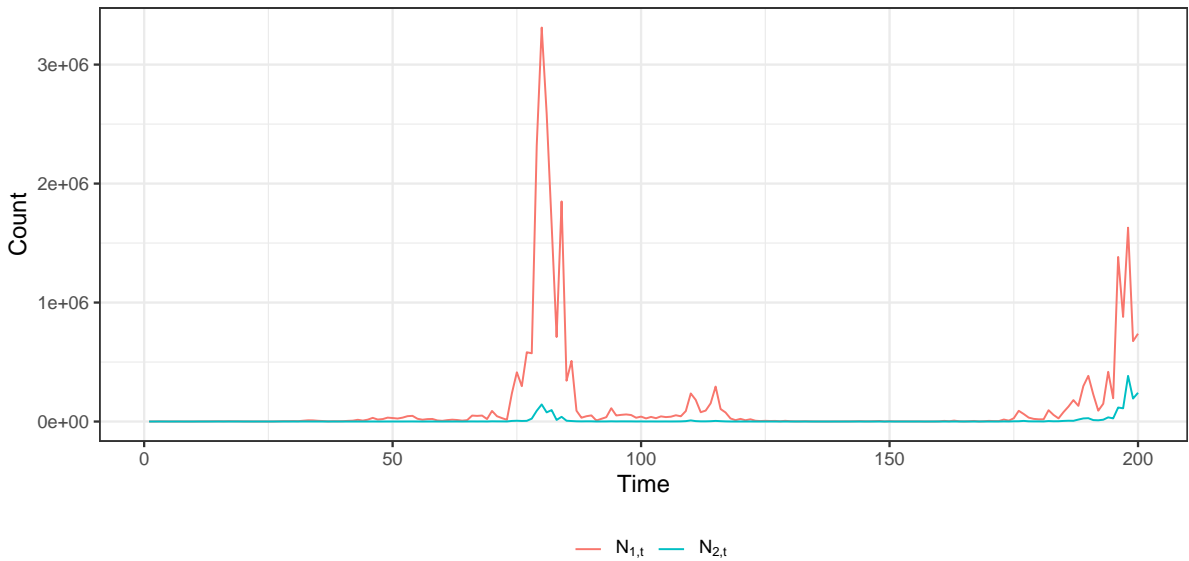


Figure 3.1: Simulated time series used in the study.

All the effects of time and delay were modeled by random walks of order 1, and a vague gamma prior mentioned in section 2.3 was considered for the precision hyperparameters. The point estimates were obtained via the posterior median, and the 95% credible intervals were obtained by taking equal probability for the tails of the posterior

distribution. To obtain the nowcasting estimates, a sample of size 1000 from the posterior predictive of  $N_1$  and  $N_2$  was considered.

Table 3.2 presents the real values, and the point and interval estimates for the fixed effects in the model, which include the intercept terms and the mean effect of the covariate. As it is presented in the table, the model can accurately recover such parameters.

Effect	True value	Posterior Median	95% Credible interval
$\alpha_1$	7.20	7.37	(6.90; 7.84)
$\alpha_2$	3.74	3.89	(3.42; 4.35)
$\nu$	-10.99	-11.54	(-14.19; -8.90)

Table 3.2: Point estimates and 95% credible intervals for fixed effects in the proposed model, with true values used in the generation of the artificial dataset.

The random effects estimates, along with the respective real values, are shown in Figure 3.2. We note that a considerable portion of the real parameters falls within the credible intervals. Moreover, the point estimates seem to track the trends in the real parameters, especially for the common time and delay effects  $\delta_t$  and  $\psi_d$ .

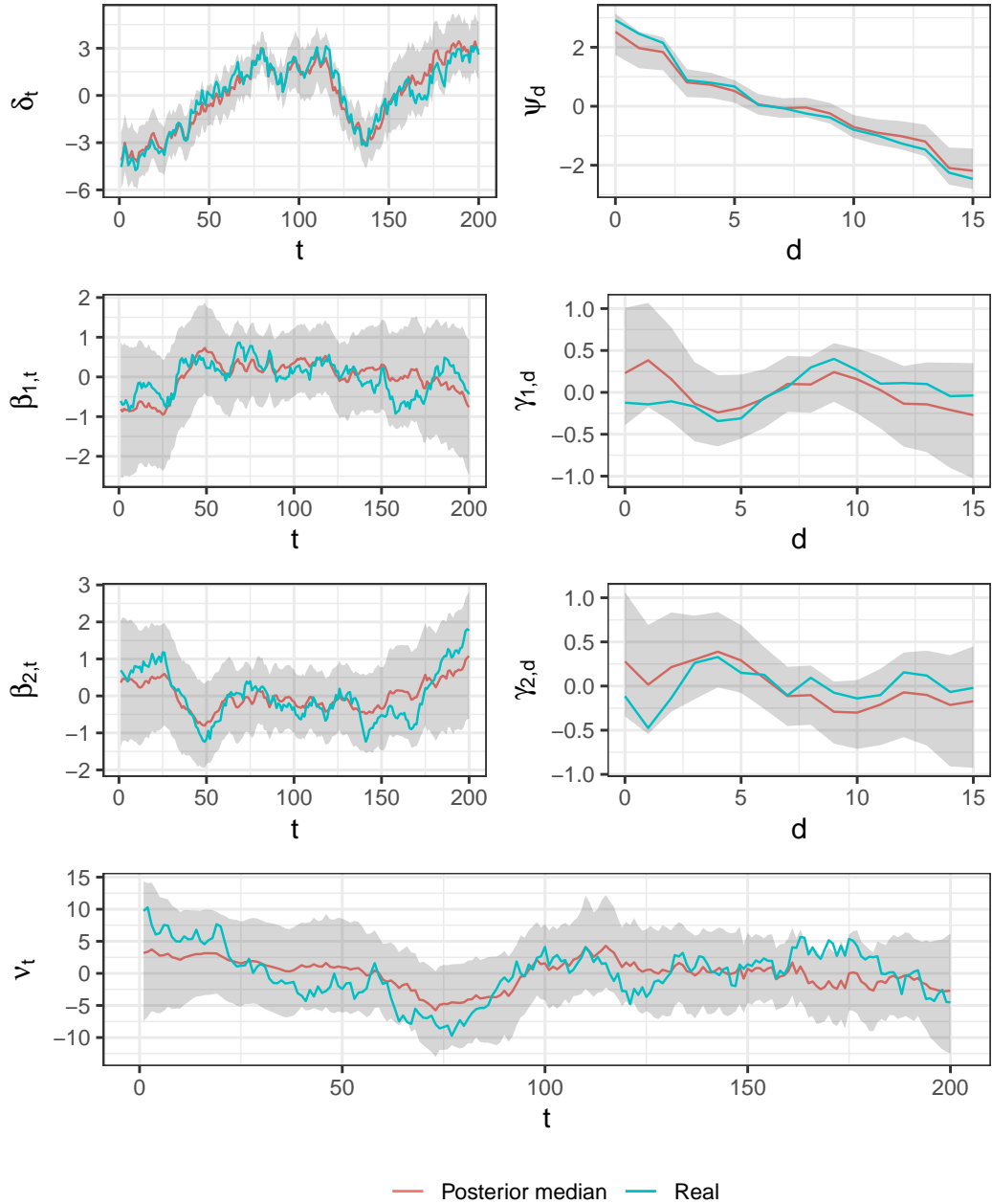


Figure 3.2: Real values, posterior median, and respective 95% credible intervals for the random effects in the proposed model.

Figure 3.3 displays the nowcasting estimates, the observed data at the moment, and the actual observed data for  $N_1$  and  $N_2$ . We observe that the corrections are notably accurate for  $N_1$ . For  $N_2$ , the corrections do not match the observed data as closely towards the end of the time series. Nonetheless, the model succeeds in correcting for the step rise in the observed data for  $N_2$  when compared to the data available at the moment, and the true values are inside the interval estimates throughout most of the time frame. The figure also gives a sense of the increase in uncertainty as the number of time

steps advances, an expected behavior since more values are unknown when computing the nowcasts for more recent values.

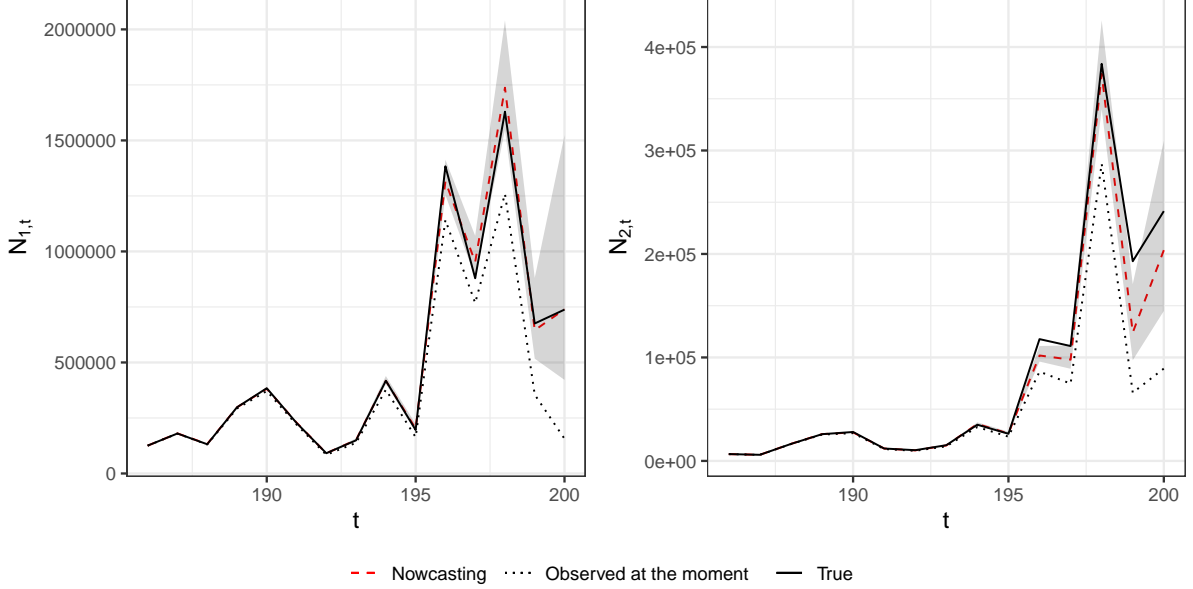


Figure 3.3: Nowcasting estimates, data observed at the moment of predictions, and true observed data artificially generated in the simulation study.

The true values and estimates for the hyperparameters are shown in Table 3.3. The model managed to capture the real values of the precisions of most of the hyperparameters. Except for an underestimation of precision, and hence overestimation of uncertainty, for the marginal delay effects  $\gamma_{1,d}$  and  $\gamma_{2,d}$ .

Hyperparameter	Real value	Posterior median	95% Credible interval
$\phi_1$	5	4.96	(4.70; 5.25)
$\phi_2$	5	5.21	(4.80; 5.66)
$\sigma_\beta^{-2}$	50	41.49	(29.71; 60.41)
$\sigma_\gamma^{-2}$	50	25.95	(14.04; 50.12)
$\sigma_\delta^{-2}$	5	4.59	(3.15; 6.47)
$\sigma_\psi^{-2}$	5	4.90	(2.66; 9.55)
$\sigma_\nu^{-2}$	1/2	0.52	(0.23; 1.67)

Table 3.3: Real values, point and interval estimates for the hyperparameters.

This study provides encouraging results with respect to the use of the proposed model. Although with less well-behaved time series than the real dataset further considered in this study and possible future practical applications to disease data, we manage to capture the real values of fixed and random effects, besides most of the hyperparameters. The nowcasting corrections have shown to be accurate and successfully recover the real data from the observed with delay.

### 3.1.2 Evaluating different models

In this subsection, we compare the true model from which the data was generated with two others in order to examine how the metrics behave and assist in model choosing. The first is a Poisson model with the same structure, i.e. same linear predictor as in Equation (2.2). The second model considered for this comparison is a model without the common terms of time and delay, which can be viewed as a proxy for using two separate univariate models for ease of workflow and implementation. It is important to mention that this model is equivalent to two univariate models except for the precision hyperparameters for the random walks of the effects of time and delay, which are assumed to be the same in this case. That is, for the independent model, the effects can be seen as two realizations of the same process, whereas using two separate models would allow for different precisions in each effect. The linear predictor in this case is given by the following:

$$\log(\lambda_{i,t,d}) = \alpha_i + \beta_{i,t} + \gamma_{i,d} + \nu_{i,t}x_t.$$

Note that, in addition to not having the common effects of time and delay, another difference between the bivariate negative binomial model fitted to this data and the independent model is the handling of the covariate coefficient. In the bivariate model used in this application, the covariate coefficient is assumed to be common to both time series, whereas in the independent model, it is considered free to vary between the two series. This latter approach is analogous to a double-univariate model, treating each time series separately.

Table 3.4 presents the results for the models evaluated, bold indicates the best results. Some metrics are also presented with respect to the last observed value, such as the last relative interval score and relative width, as a way to obtain insights regarding the estimation under most uncertainty. It is notable that the Poisson model performs worse than the other two in most cases, presenting smaller intervals but also smaller coverage. Both negative binomial models perform equivalently in terms of coverage, with the bivariate model exhibiting smaller relative widths and a larger mean interval score. However, when examining the last relative interval score and the metrics for the posterior predictive samples, the proposed model performs better, with significant differences observed in the mean variogram score and the last energy and variogram scores. While the independent model performs better in terms of the mean energy score, its performance is comparable to the bivariate negative binomial model. In this case, the proposed model would be a suitable choice.



	Bivariate (NB)	Bivariate (Poisson)	Independent
MAPE ( $N_1$ )	<b>2.156</b>	4.246	3.805
MAPE ( $N_2$ )	5.862	6.044	<b>4.602</b>
Coverage ( $N_1$ )	<b>0.867</b>	0.800	<b>0.867</b>
Coverage ( $N_2$ )	<b>0.800</b>	0.467	<b>0.800</b>
Coverage (both)	<b>0.667</b>	0.333	<b>0.667</b>
Last coverage (both)	1.000	0.000	1.000
mean relWidth	0.344	<b>0.137</b>	0.367
mean relIS	0.780	2.265	<b>0.648</b>
last relWidth	2.137	<b>0.548</b>	2.281
last relIS	<b>2.137</b>	8.094	2.281
mean Energy Score	19873.920	30157.008	<b>19692.964</b>
mean Variogram Score	<b>1282.004</b>	5664.418	2660.245
last Energy Score	<b>66122.889</b>	219450.046	92603.079
last Variogram Score	<b>4710.258</b>	72361.541	31261.713

Table 3.4: Performance results for models considered.

As expected, the metrics point to the true model used to generate the data, the bivariate Negative Binomial model with common and marginal effects. For more detail on the nowcasting estimates for each model, Figures A.2 and A.3 in the Appendix show the estimates according to the Poisson and Independent models, respectively.

## 3.2 Dengue and chikungunya in Rio de Janeiro

In this section, we present an application of the proposed model to weekly data of dengue and chikungunya in the state of Rio de Janeiro from 2017 to 2019, shown in Figure 1.1. The data consists of suspected cases of dengue and chikungunya in the city of Rio de Janeiro reported in the Sistema de Informação de Agravos de Notificação (SINAN). The aggregated data are available at InfoDengue.

This section is divided into two parts. In subsection 3.2.1 we explore the dataset, motivating the choice of the maximum possible delay, and conducting a sliding window analysis evaluating a simpler formulation of the model with different window sizes to assess the performance. In subsection 3.2.2 we apply different structures of the model using the number of tweets with mention of the term “dengue” as a covariate. Four models are considered with different formulations for the linear predictor, including or excluding some of the effects. The application takes into account the sliding window size chosen from subsection 3.2.1.

### 3.2.1 Sliding windows analysis

The mean time between the onset of symptoms and reporting was 5.5 and 4.3 weeks, for chikungunya and dengue, respectively. At the same time, the median time is 3 and 2 weeks, respectively, for dengue and chikungunya. Although the mean and median delay

might seem low, the difference between these measures in both cases can give us a sense of the asymmetry in this case. In fact, fewer than 80% of cases are reported within 5 weeks of symptoms onset.

Table 3.5 shows the cumulative proportion of reported cases within 5, 10, 15, and 20 weeks. We observe that over 90% (and close to 95%) of cases are reported within 15 weeks. Figure 3.4 shows the non-cumulative proportions of cases reported by the number of weeks of delay along with a 1% threshold dashed line, note that after 15 weeks both dengue and chikungunya present less than 1% of weekly reported cases. Consequently, we have decided to use a maximum possible delay of 15 weeks.

Delay (weeks)	Dengue (%)	Chikungunya (%)
5	76.40	67.25
10	89.35	86.01
15	94.87	93.46
20	97.46	96.59

Table 3.5: Cumulative percentage of reported cases for dengue and chikungunya by number of weeks of delay.

We aim to nowcast the number of cases in a given week for dengue and chikungunya. Hence, the proposed model with  $p = 2$  will be applied. Our first approach is to apply a simpler formulation of the model in Equation (2.2), without the covariate term, considering sliding windows of different sizes.

In order to evaluate the capacity of the model to recover the observable counts in different stages of an epidemic – in fact, we are especially interested in the ability to perform accurate corrections in moments of an increasing number of cases – we performed the sliding windows analysis from March 27th to September 8th, 2019 (epidemiological weeks 13 to 37 from 2019), since this period contemplates the rise and fall of the number of cases for both dengue and chikungunya. This behavior can be observed in Figure 1.1.

All models presented throughout this section consider a random walk of order 2 for the time effects, explicit in Equation (3.2), and a random walk of order 1 for delay effects and the covariate’s coefficient. We considered the same vague gamma prior from section 2.3 for the precision hyperparameters  $1/\sigma^2$  in Equations (2.4) to (2.8) and the size of the Negative Binomial distribution  $\phi_i$ .

$$\beta_t \mid (\beta_{t-1}, \beta_{t-2}, \sigma_\beta^2) \sim N(2\beta_{t-1} - \beta_{t-2}, \sigma_\beta^2) \quad (3.2)$$

After calculating the approximation, a sample of size 1000 was taken from the joint posterior in the inference procedure. The point estimations are given by the median of the posterior predictive distribution, and we took the 95% credible intervals considering the same probability for the tails of the posterior predictive for the interval estimation.

Table 3.6 displays the window sizes, as well as the time required to evaluate the models and performance metrics. These experiments were performed on a computer running a 64-bit Windows 10 operating system, equipped with an Intel i5-8265S CPU and 12 GB of RAM. In addition to the mean of metrics across all the correction windows, we also

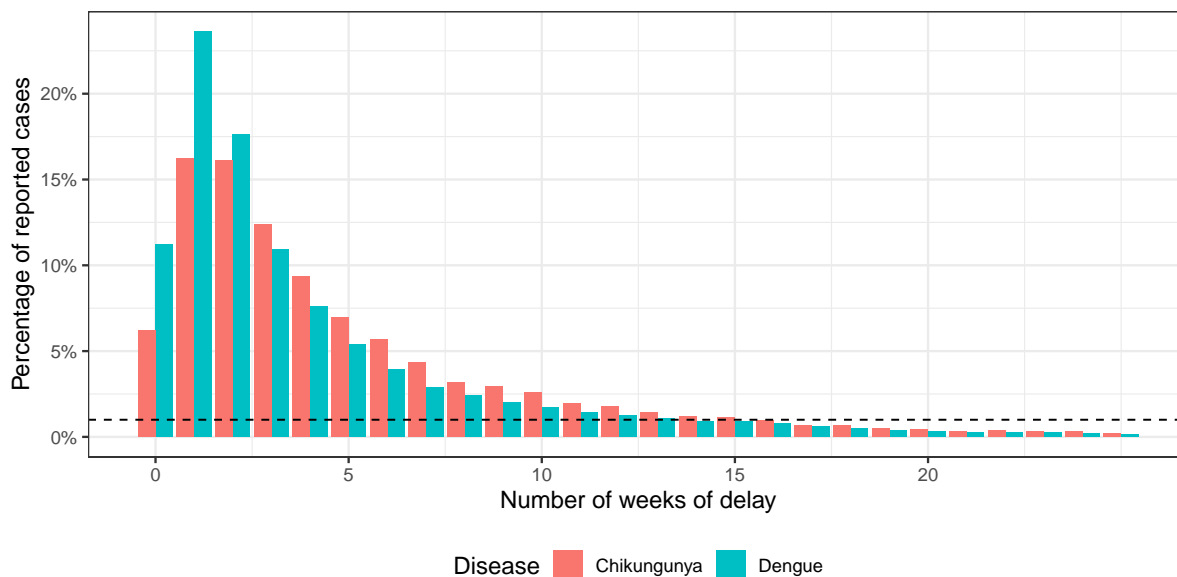


Figure 3.4: Proportion of reported cases of dengue (Blue) and chikungunya (Red) by number of weeks of delay along with 1% of reported cases threshold (dashed line).

present some metrics averaged over the last predictions made for each window. Assessing these metrics is particularly insightful because they concern the predictions under the most uncertainty.

The results indicate that increasing the window size leads to better performance, except for the time taken to run the models. However, a 70 weeks window size provides comparable results to using the full data, which is equivalent to a window size of 116, despite the latter taking twice as long to run. Moreover, we observe that as the window size increases, the relative interval widths also increase, while the relative interval scores decrease. This stands to reason, given that wider intervals (usually) result in better coverage and smaller penalty terms in the interval scores. The metrics averaged over the last predictions yield similar results. Mean energy and variogram scores also point to bigger windows with comparable results for a window of size 70 and the full data.

Given the results in Table 3.6 we can proceed to a more detailed analysis with a window of size 70. Figure 3.5 displays the results of the corrections of chikungunya cases for each window. Note that, in most cases, the proposed model is able to recover the count, estimating its value close to the actual observed data several weeks later. This happens especially during periods of increasing and peak numbers of cases, which is highly relevant in a nowcasting problem. For moments a few weeks later the peak, the model is unable to capture the actual height of the peak accurately, but it still succeeds in estimating the current number of cases. This is evident in windows 16 and 17, for example. The corrections for dengue are displayed in Figure B.1 in Appendix B and show similar results.

	Window Size (weeks)					
	30	40	50	60	70	Full data
Fit time (in minutes)	<b>2.410</b>	3.540	4.730	6.480	8.100	17.830
MAPE (Chikungunya)	11.479	10.685	9.662	8.868	8.191	<b>7.753</b>
MAPE (Dengue)	9.589	9.847	9.228	8.228	<b>7.625</b>	8.052
Coverage (Chikungunya)	0.237	0.351	0.433	0.516	0.569	<b>0.612</b>
Coverage (Dengue)	0.258	0.254	0.307	0.374	<b>0.426</b>	0.378
Coverage (both)	0.192	0.222	0.283	0.358	<b>0.404</b>	0.373
Last coverage (both)	0.840	0.960	0.960	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
mean relWidth	<b>0.336</b>	0.364	0.404	0.453	0.497	0.526
mean relIS	4.279	3.890	3.264	2.586	<b>2.169</b>	2.202
mean last relWidth	<b>1.685</b>	1.793	1.910	2.130	2.347	2.528
mean last relIS	2.223	<b>1.898</b>	1.956	2.130	2.347	2.528
mean Energy Score	16775.018	16778.206	16768.821	16744.975	16732.822	<b>16731.638</b>
mean Variogram Score	19942.819	19664.789	19423.708	19193.965	<b>19022.499</b>	19082.418

Table 3.6: Window sizes considered, time taken to fit the model, and respective performance metrics, with optimum values in bold.

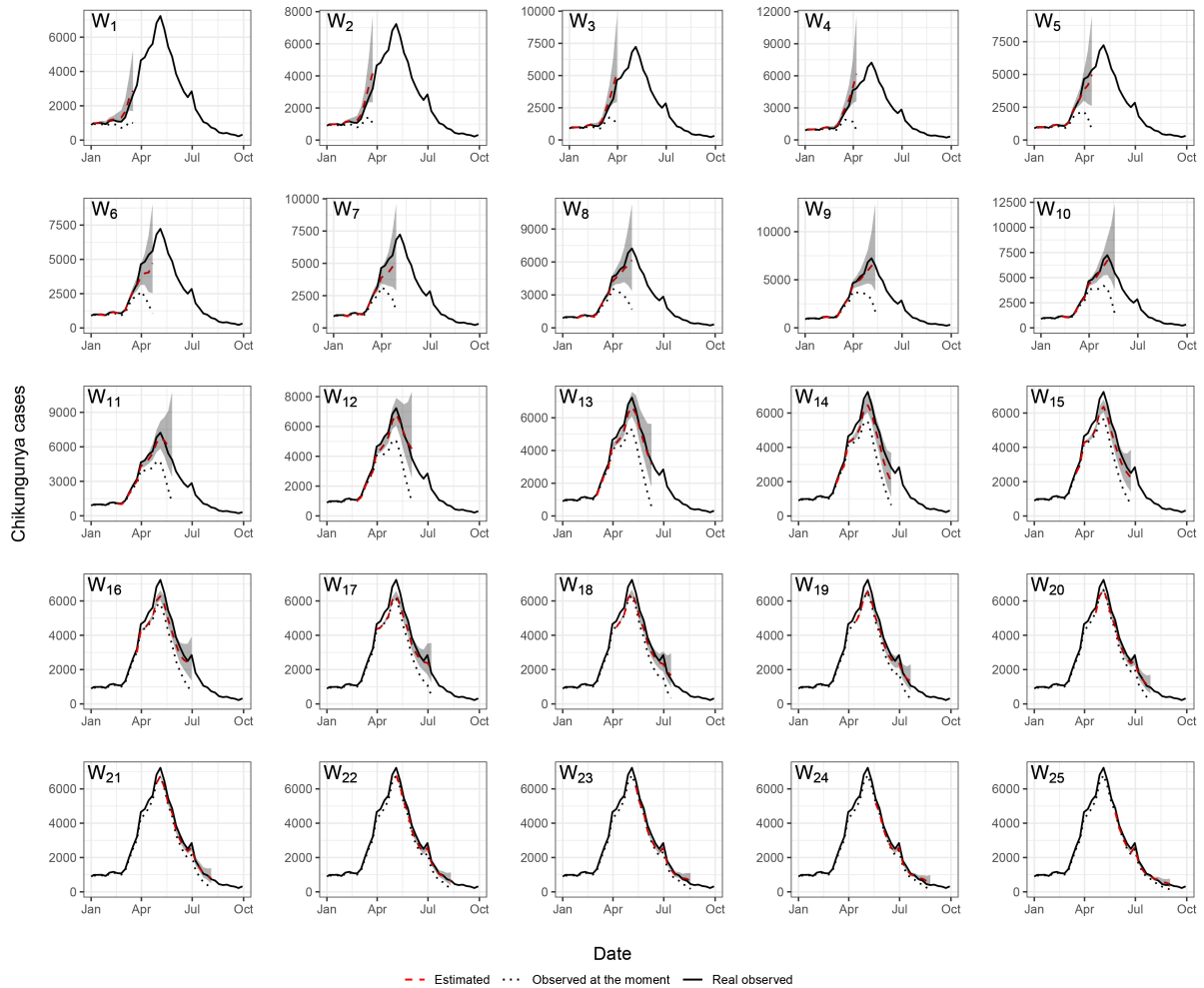


Figure 3.5: Nowcasting for each of the 25 windows of 70 weeks to which the model was fitted.

In the next subsection, we compare a few formulations of the model with sliding windows of size 70 weeks and evaluate the results.

### 3.2.2 Comparing model structures

In this subsection, we aim to compare different model structures using a sliding window of size 70 weeks, as determined in subsection 3.2.1. We include the number of tweets mentioning the word “dengue” located in Rio de Janeiro, available from the InfoDengue system, as a covariate. Figure 3.6 shows the time series of this covariate. We notice that a dynamic coefficient may be suitable as there appears to be a similar trend between the series of tweets and the number of cases of dengue and chikungunya in 2019. However, the same similarity does not happen during the epidemic moment of 2018, which may indicate that the effect of the number of tweets becomes more significant as we approach 2019.

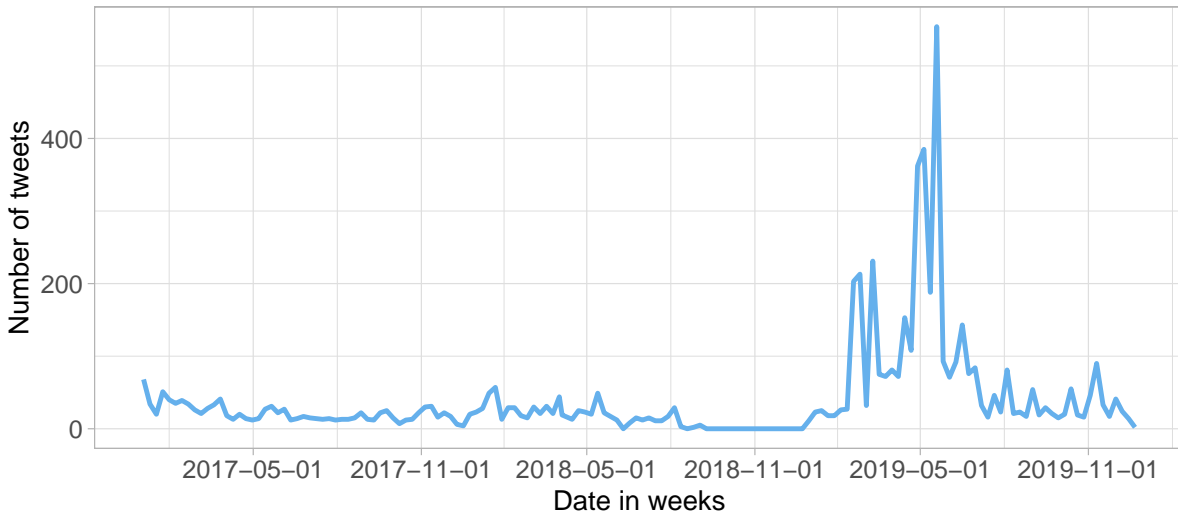


Figure 3.6: Number of tweets published in Rio de Janeiro mentioning the term “dengue”. From 2017 to 2019.

Table 3.7 displays the models considered for the present application, along with their corresponding linear predictor structures. To account for the possibility of using two separate univariate models, we include a model without the common terms as a proxy, as described in the previous section. A comparison between results when using two univariate models and the proxy bivariate model used here is presented in Figures C.1 and C.2 in Appendix C. Additionally, we consider the complete proposed model, as well as two specific cases: the proposed model with a fixed coefficient for the covariate and the model without the covariate. It is worth noting that all models in the Table are particular cases of M1.

Model description	Linear predictor ( $\log(\lambda_{i,t,d})$ )
M0 - Independent model (model without common effects)	$\alpha_i + \beta_{i,t} + \gamma_{i,d}$
M1 - Complete model	$\alpha_i + \beta_{i,t} + \gamma_{i,d} + \delta_t + \psi_d + \nu_{i,t}x_t$
M2 - Complete model with static covariate effect	$\alpha_i + \beta_{i,t} + \gamma_{i,d} + \delta_t + \psi_d + \nu_i x_t$
M3 - Model without the covariate	$\alpha_i + \beta_{i,t} + \gamma_{i,d} + \delta_t + \psi_d$

Table 3.7: Models considered in the application and respective linear predictor structures.

The performance results for each model are presented in Table 3.8. It is observed that M2 performs better when considering the interval metrics, and is the second-best model when considering both MAPE, with an error difference of less than 0.1% to M0. The baseline M0 performs slightly better when examining MAPE, variogram, and energy scores. Despite this, M0 is the worst of the four models when considering interval metrics.

	M0	M1	M2	M3
MAPE (Chikungunya)	<b>8.042</b>	8.102	8.090	8.191
MAPE (Dengue)	<b>7.419</b>	7.551	7.555	7.625
Coverage (both)	0.392	0.401	0.400	<b>0.404</b>
mean relWidth	0.517	0.500	<b>0.497</b>	<b>0.497</b>
mean relIS	2.188	2.159	<b>2.153</b>	2.169
mean last relWidth	2.482	2.344	<b>2.313</b>	2.347
mean last relIS	2.482	2.344	<b>2.313</b>	2.347
mean Energy Score	<b>16731.6</b>	16731.8	16733.7	16732.8
mean Variogram Score	<b>18950.2</b>	19028.0	19031.0	19022.5

Table 3.8: Performance metrics for models used in the application.

Although the differences in model performance may seem small when comparing the aggregate results, some conclusions become clearer when examining certain metrics across the sliding windows as shown in Figure 3.7. Notably, models with the common terms present smaller intervals without sacrificing the interval score, which means the coverage is maintained. Concerning the point estimates, the results are not as clear as for the interval metrics. However, we can see that M0 yields larger errors for chikungunya in moments of high number of cases such as windows 10 to 12. In contrast, M0 shows similar performance for dengue corrections when compared to the bivariate models.

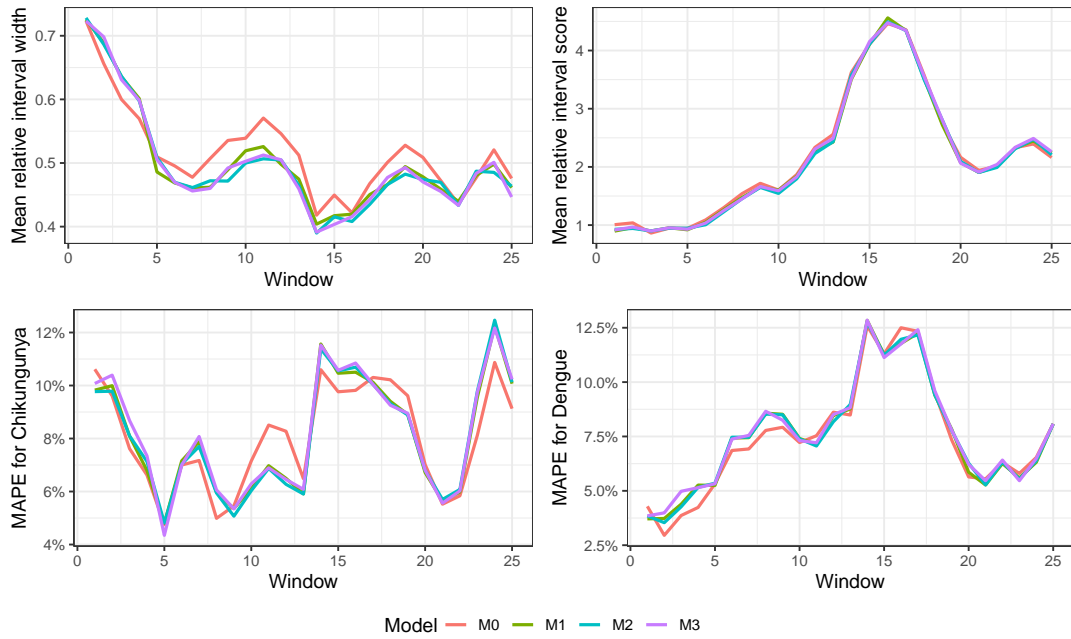


Figure 3.7: Relative interval width, relative interval score, and MAPE for each of the 25 sliding windows colored by model.

In order to investigate the time and delay structure estimates, we chose to fit M2 to the full data after observing the results in Table 3.8 and Figure 3.7. The decision to use the full data was made to obtain estimates of time effects for all the time steps and compare them to the original time series. The results are presented in Figure 3.8. Notably, the time effects follow the actual behavior of both original time series in Figure 1.1. As for the delay, the effects exhibit a similar pattern to the proportion bars presented in Figure 3.4, with a higher proportion of reports for dengue at the beginning, and an inversion as the amount of delay increases.

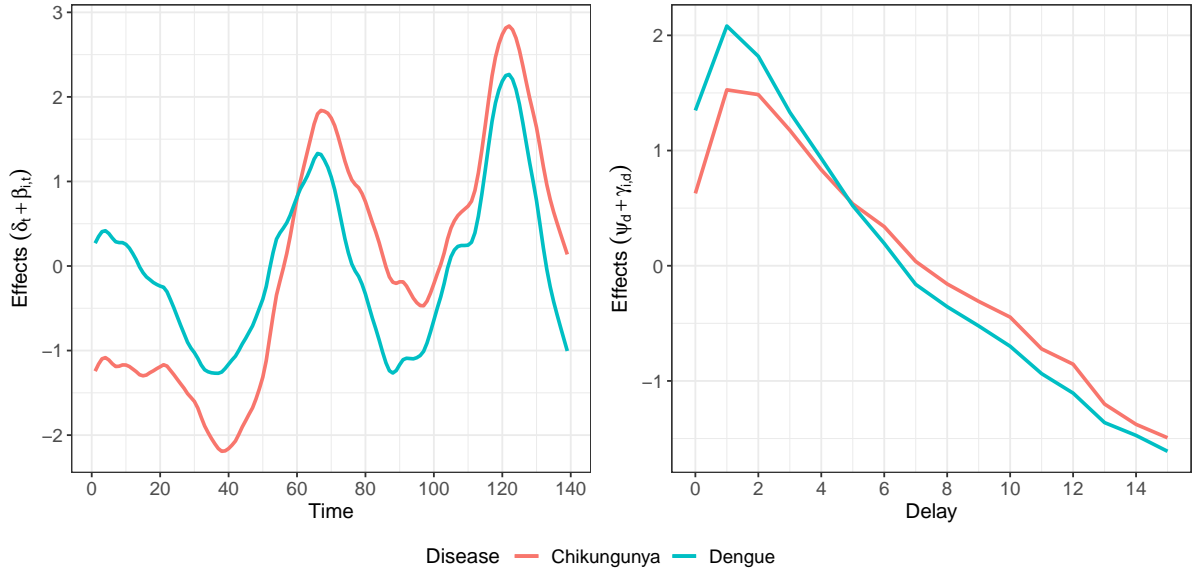


Figure 3.8: Effects of time and delay for dengue (blue) and chikungunya (red).

Table 3.9 displays the intercept terms  $\alpha_1$  and  $\alpha_2$ , as well as the fixed coefficients for the number of tweets  $\nu_1$  and  $\nu_2$ . We observe a higher intercept term for the chikungunya cases, which stands to reason since the number of chikungunya cases is higher throughout most of the analyzed period. Although the coefficients for the covariate effect do not significantly differ from 0 based on the interval estimates, we retained them in the model.

Effect	Median	95% credible interval
$\alpha_1$ (Chikungunya)	2.77	( 2.63; 2.92)
$\alpha_2$ (Dengue)	2.46	( 2.32; 2.61)
$\nu_1$ (Chikungunya)	0.02	(-0.02; 0.07)
$\nu_2$ (Dengue)	0.01	(-0.03; 0.06)

Table 3.9: Fixed effects estimates for M2 with full data.

The posterior estimates of  $\phi_i$  and precision hyperparameters for the same model are presented in Table 3.10. The smaller precision for the correlation-inducing parameters,  $\delta_d$  and  $\psi_d$ , when compared to the  $\beta$  and *gamma* terms, show that these have more importance in the linear predictor and the model is not close to a double univariate approach.



Hyperparameter	Median	95% credible interval
$\phi_1$	3.66	(3.35; 3.99)
$\phi_2$	4.43	(4.04; 4.87)
$\sigma_\beta^{-2}$	2848.90	(1180.54; 5824.83)
$\sigma_\gamma^{-2}$	104.59	(43.40; 256.19)
$\sigma_\delta^{-2}$	287.10	(151.71; 532.99)
$\sigma_\psi^{-2}$	9.56	(3.96; 18.94)

Table 3.10: Posterior estimates of hyperparameters for M2 with full data.

The low values for  $\phi_i$ ,  $i = 1, 2$ , indicate that the Poisson distribution is likely not fit for this application. To examine this, we performed nowcasting with a Poisson distribution and compared it with the Negative Binomial approach for one of the windows considered (Window 11). The corrections are displayed in Figure 3.9. We notice that although the Poisson model performs better for chikungunya correction, it confidently underestimates the number of dengue cases, yielding point estimates that fall below the true value and credible intervals that do not contain it.

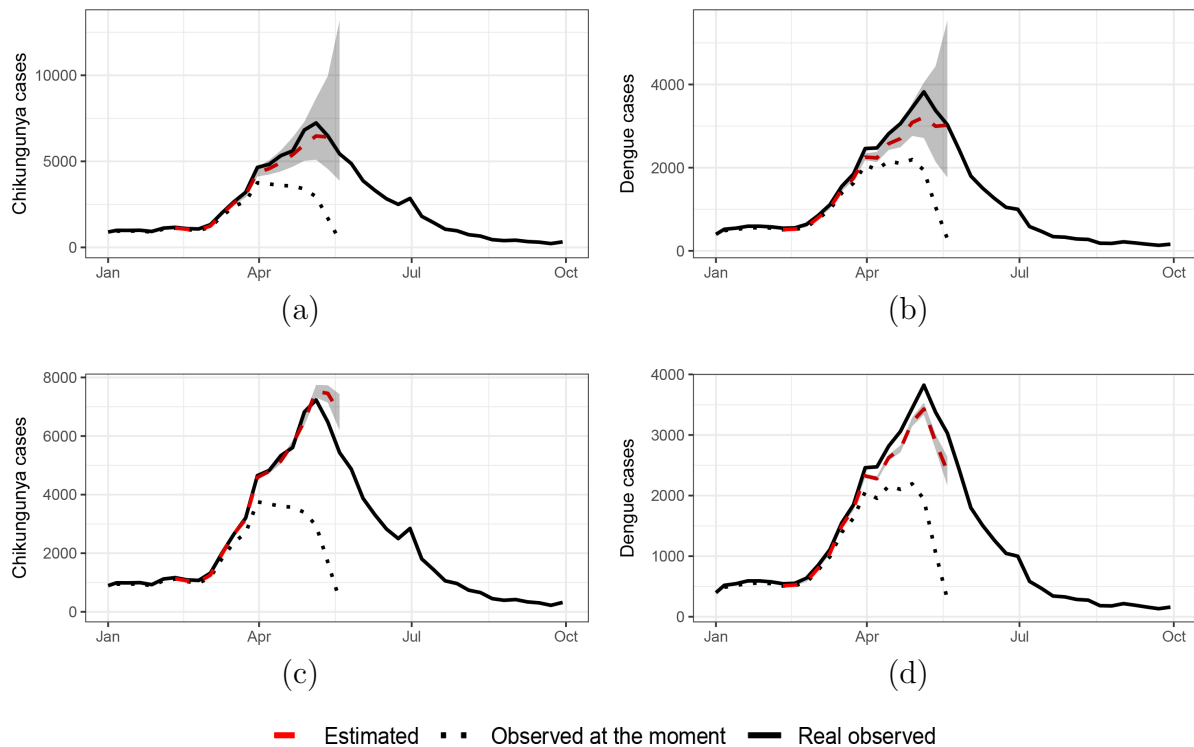


Figure 3.9: Nowcasting using the Negative Binomial (panels (a) and (b)), and the Poisson (panels (c) and (d)) distributions considering full data until May 19th, 2019.

### 3.2.3 Prior sensitivity analysis

Since we are using a different prior from the default INLA Gamma prior for the precision and  $\phi_i$  hyperparameters, it is of interest to assess the sensitivity of the model to

the prior choice and its implications in the final nowcasting estimates. We consider some prior distributions suggested for hierarchical models from Gelman (2006) implemented for use as priors for precision hyperparameters in INLA by Gómez-Rubio (2020). The priors examined in the study are shown in Table 3.11.

Default Gamma( $1, 5 \times 10^{-5}$ )
Gamma(0.001, 0.001) used in this work
Half-Normal with precision $\tau_0 = 0.001$
Half-Cauchy with scale parameter $\gamma = 25$
Half-t with 3 degrees of freedom
Improper Uniform prior

Table 3.11: Priors considered for a sensitivity analysis

For each of the priors, we applied the model M2 to the same data used to obtain the estimates in Figure 3.8 and Table 3.10, covering 140 weeks of data from January 1st, 2017 to September 8th, 2019. The resulting posterior marginals for each of the hyperparameters  $\sigma^{-2}$  according to the priors of choice are displayed in Figure 3.10. We notice that, for the overdispersion hyperparameters  $\phi_i$ , the prior choice does not seem to influence the posterior.

In contrast, the posterior distributions for the precisions of time and delay effects present different behaviors according to the priors of choice. The Half-Normal, Half-t, Half-Cauchy, and Uniform priors produce similar results in this case. Whereas both Gamma priors present different results depending on the hyperparameter. For Instance,  $\sigma_{\beta}^{-2}$ , the precision for the  $\beta_{t,i}$  effects, we notice that the applied Gamma prior with parameters  $\alpha = \beta = 0.001$  is more concentrated on smaller values while the default prior has an opposite result.

Regarding the other three precision hyperparameters, using the default Gamma prior yields posterior distributions with higher probabilities for larger values of the hyperparameter. Additionally, choosing the alternative (applied) Gamma prior produces results that are generally more similar to the other four priors, occasionally still very similar to the results obtained with the default Gamma prior. An instance of this is observed in the case of  $\sigma_{\phi}^{-2}$ .

The median posterior and interval estimates for the hyperparameters according to each prior can be found in Table D.1 in the Appendix D.

Since there are differences in the posterior for certain hyperparameters, it is crucial to analyze how the choice of priors affects the final nowcasting estimates. Figure 3.11 presents the point and interval estimates according to the priors and respective interval widths. Although there are slight variations between the predictions and interval widths, they do not appear to be influenced by the choice of prior.

In order to investigate whether this behavior is due to the fact that the moment being nowcasted has a decreasing number of cases, we fitted the model aiming to nowcasting a moment with a high number of cases, the same data used in Figure 3.9. The results, shown in Figure 3.12, are similar, presenting mild differences for point and interval estimates according to the priors. The small fluctuations of interval width and point estimates

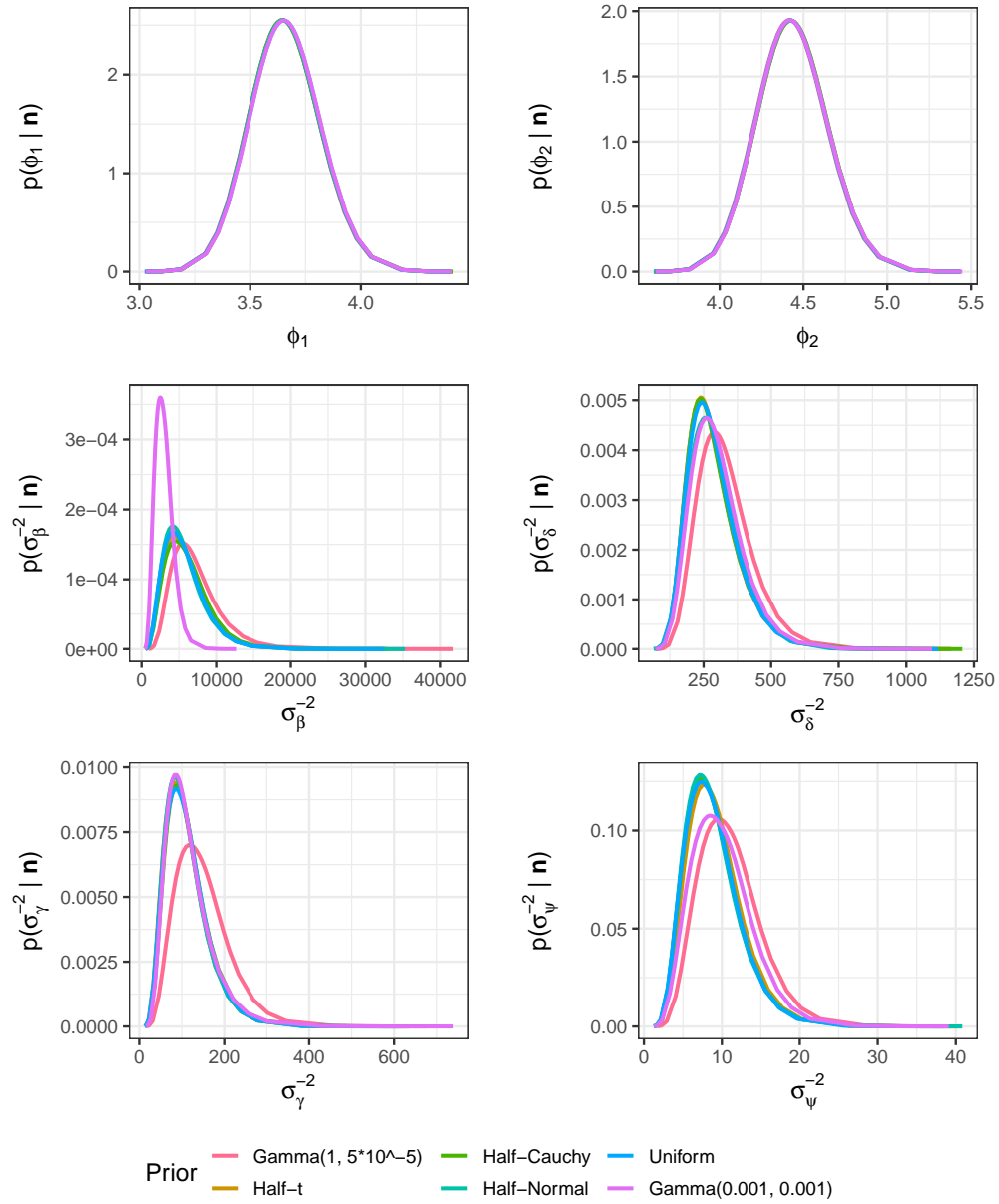


Figure 3.10: Posterior marginals for the hyperparameters according to prior of choice.

for more recent nowcasted steps may also be due to the fact that we are sampling from the predictive posterior distribution of the true observable counts and there is more uncertainty associated with the last steps. See Table 2.1, where the  $T$ -th row is almost entirely not-yet-observed.

Hence, the nowcasting estimates have shown to be robust to the choice of prior distribution for the precision and  $\phi_i$  hyperparameters.

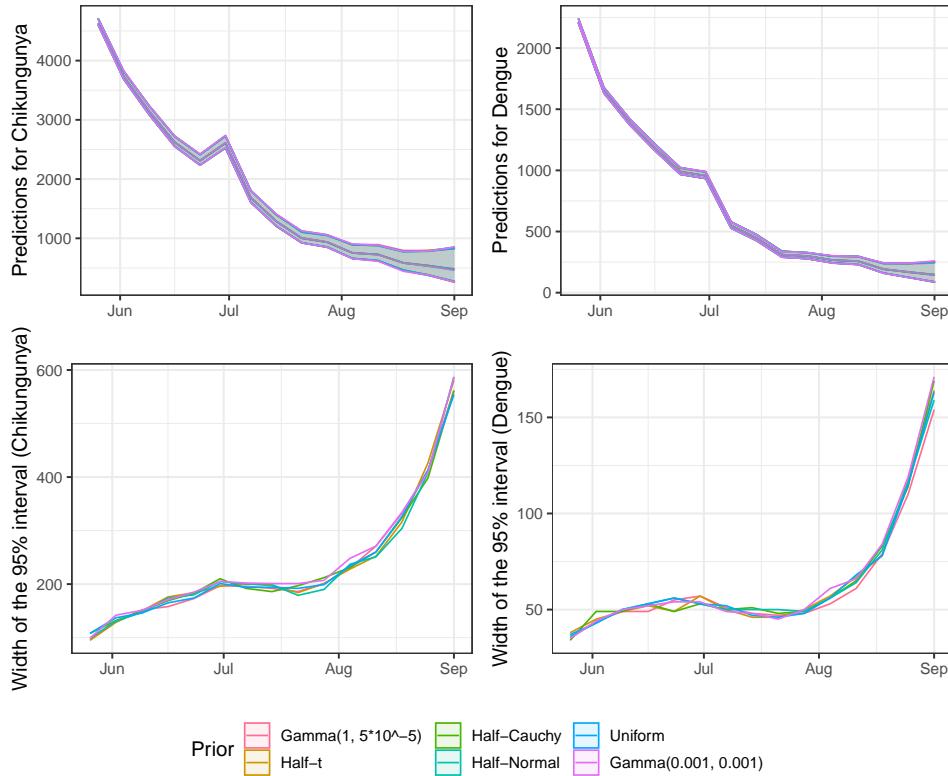


Figure 3.11: Nowcasting estimates and interval widths according to prior of choice.

### 3.3 Discussion

The study with the artificial dataset demonstrated that the proposed model is capable of accurately capturing the fixed and random effects, as well as the associated hyperparameters, despite the common delay effect being generated according to a different process than the one used to fit the model and despite the difference in scale between the generated time series.

When comparing different models on the artificial data, the metrics pointed to the true model used to generate the data, although an equivalence was found for the interval coverages when compared to the model without the common effects of time and delay. The use of the Poisson distribution, in this case, has been shown to underestimate the uncertainty of the nowcasting corrections and provide very short intervals that rarely contain the real value.

The application in real data for dengue and chikungunya also provided encouraging results. We managed to find a suitable maximum acceptable delay for the dataset and proceed with a sliding windows approach to find a smaller window size that provides equivalent results to the use of the entire dataset. For the sliding windows corrections, we observed that the model was able to capture the real current situation of the number of cases in the different moments of the epidemic curve, which is one of the main features of interest in a nowcasting model.

Different model structures were compared using the number of tweets as a covariate,

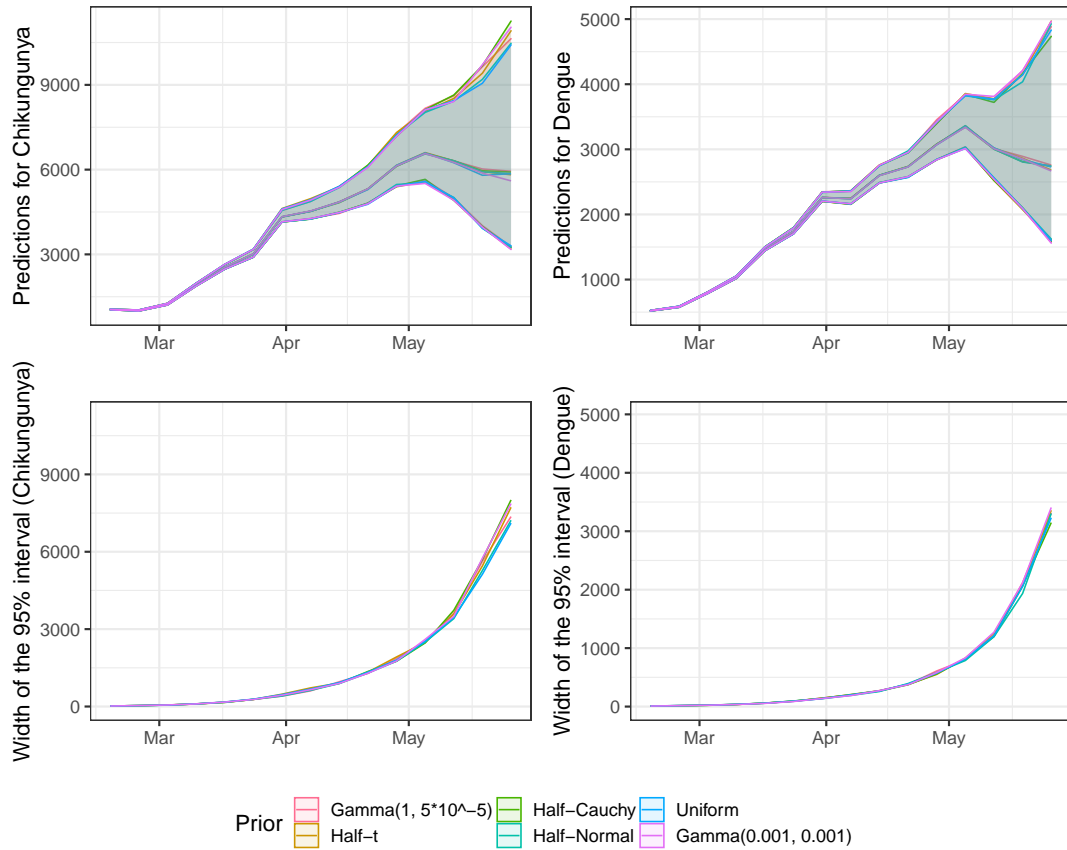


Figure 3.12: Nowcasting estimates and interval widths for a moment of high number of cases according to prior of choice.

and the model with the fixed covariate effect showed more suitable for this application. It is also noteworthy that every model with common terms presented better performance than the approach without the common effects, providing smaller intervals while maintaining coverage.

These smaller intervals obtained with a bivariate approach can facilitate a possible future short-term forecasting, once we would have a smaller uncertainty propagated to the prediction of future steps when compared to univariate approaches, hence having more reasonable interval estimates in this case.

In an application to the full data, we note that the time effects estimates accurately depict the relation between both series of interest, evolving similarly to the previously shown real data. The same goes for the delay structure, showing similar behavior to the weekly proportions of reported cases by amount of delay.

We also examined an application with the Poisson distribution and found that the results did not encourage its use in this particular application. Some corrections underestimated the number of cases with very little uncertainty when compared to the negative binomial approach.

To assess the robustness of the model to prior distribution choices for the hyperparameters, we performed a sensitivity analysis. The results indicate that the posterior

distributions of the hyperparameters may vary depending on the prior distribution, even though we are using vague priors. Despite these variations, the final nowcasting estimates do not appear to be affected by this variation, as both the estimates and the quality of the intervals exhibit only a slight variation across the different prior distributions.

# Chapter 4

## Concluding remarks

We managed to propose a Bayesian hierarchical multivariate framework for correcting reporting delays in disease surveillance data. The model and implementation present practical feasibility and yet the flexibility to accommodate different effects of time and delay for both time series and also covariates with effects varying over time. The model can be used with any data presented in the structure of Table 2.1, where an estimation for the run-off triangle is needed.

Covariate effects can be equal or different for both series, and the user may also choose to use only the fixed effect of the covariate, without the dynamic effect. The model also accepts a different distribution for count data and, depending on the application, a Poisson distribution can be examined. Especially if the  $\phi_i$  hyperparameters estimation is a high value, having in mind that these hyperparameters can be seen as the inverse of the overdispersion.

Along the work, we considered metrics to jointly evaluate multivariate predictions, which made it possible to simplify model evaluation for the choice of window size and model structure. In particular, the relative interval width and interval score made it easier to assess the quality of interval predictions without having to examine a different metric for each prediction. Additionally, the energy and variogram scores were useful in a similar manner. Nevertheless, the problem of multivariate scoring rules for probabilistic forecasting is one with room for improvement in literature.

We applied the model to artificial and real data and obtained encouraging results for the use of a multivariate approach. When compared to a univariate-equivalent approach, the model generated smaller intervals without sacrificing coverage of the real values, while producing similar point estimates. This relationship was observed with or without the covariate and indicates the potential for more accurate forecasting. Typically, intervals for future predictions can quickly become unreasonably large due to the propagated uncertainty, but smaller intervals could provide more reliable forecasts for at least a few weeks.

The ability to rapidly obtain estimates and samples from the posterior with INLA facilitated a sliding windows approach to identify an “optimal” window size for the application. This is also important for the possible future development of an R package and implementation of the model in a surveillance system.

Other covariates apart from the number of tweets can be investigated in future appli-

cations. The index of google searches containing terms related to dengue or chikungunya, obtainable from google trends, is a possibility that has been shown useful for nowcasting problems in work such as Miller et al. (2022), though only with fixed covariate effects. Environmental variables such as temperature, rain, and humidity are also of possible use if available.

In future work, other formulations can be considered. For example, it would be interesting to explore the incorporation of spatial correlation into the model or to allow the delay to evolve over time with an interaction term as it is thought that delays might be longer in moments of high number of infections.



# Bibliography

- Alves, M. B., Migon, H. S., Marotta, R., and Santos Jr, S. V. (2022). k-parametric dynamic generalized linear models: a sequential approach via information geometry. *arXiv preprint arXiv:2201.05387*.
- Barbosa, M. T. S. and Struchiner, C. J. (2002). The estimated magnitude of aids in brazil: a delay correction applied to cases with lost dates. *Cadernos de Saúde Pública*, 18:279–285.
- Bastos, L. S., Economou, T., Gomes, M. F., Villela, D. A., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., and Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in medicine*, 38(22):4363–4377.
- Bastos, L. S., Niquini, R. P., Lana, R. M., Villela, D. A., Cruz, O. G., Coelho, F. C., Codeço, C. T., and Gomes, M. F. (2020). Covid-19 e hospitalizações por srag no brasil: uma comparação até a 12<sup>a</sup> semana epidemiológica de 2020. *Cadernos de Saúde Pública*, 36.
- Berry, L. R. and West, M. (2020). Bayesian forecasting of many count-valued time series. *Journal of Business & Economic Statistics*, 38(4):872–887.
- Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618.
- England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3):443–518.
- Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.
- Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2021). Nowcasting the covid-19 pandemic in bavaria. *Biometrical Journal*, 63(3):490–502.

- Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the stec o104: H4 outbreak in germany, 2011. *Biometrics*, 70(4):993–1002.
- Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90(12):1–37.
- Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin: The Journal of the IAA*, 23(2):213–225.
- McGough, S. F., Johansson, M. A., Lipsitch, M., and Menzies, N. A. (2020). Nowcasting by bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS computational biology*, 16(4):e1007735.
- Miller, S., Preis, T., Mizzi, G., Bastos, L. S., Gomes, M. F. d. C., Coelho, F. C., Codeço, C. T., and Moat, H. S. (2022). Faster indicators of chikungunya incidence using google searches. *PLOS Neglected Tropical Diseases*, 16:1–16.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Renshaw, A. E. and Verrall, R. J. (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4(4):903–923.
- Rotejanaprasert, C., Ekapirat, N., Areechokchai, D., and Maude, R. J. (2020). Bayesian spatiotemporal modeling with sliding windows to correct reporting delays for real-time dengue surveillance in thailand. *International Journal of Health Geographics*, 19(1):1–13.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.
- Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421.
- Salmon, M., Schumacher, D., Stark, K., and Höhle, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal*, 57(6):1051–1067.
- Scheuerer, M. and Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334.
- Stoner, O. and Economou, T. (2020). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, 76(3):789–798.

# Appendix A

## Supplementary material for the study with artificial dataset

In this section we present additional material for the study with the artificial dataset. Figure A.1 displays the standardized version of the time series shown in Figure 3.1.

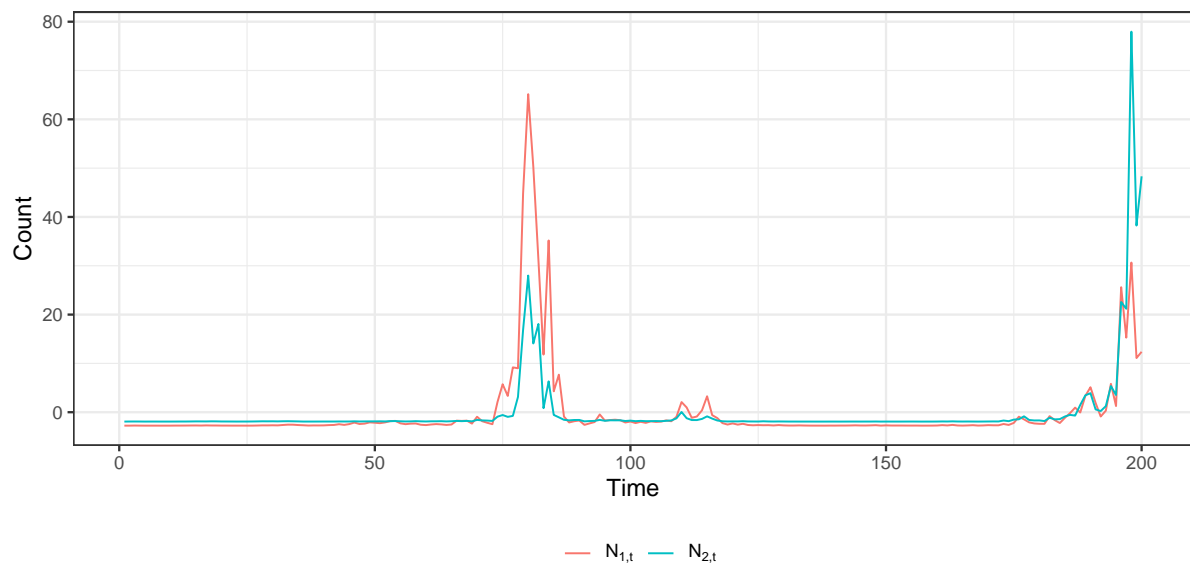


Figure A.1: Standardized version of simulated data in Figure 3.1

Nowcasting results for the Poisson and independent models are presented in Figures A.2 and A.3, respectively.

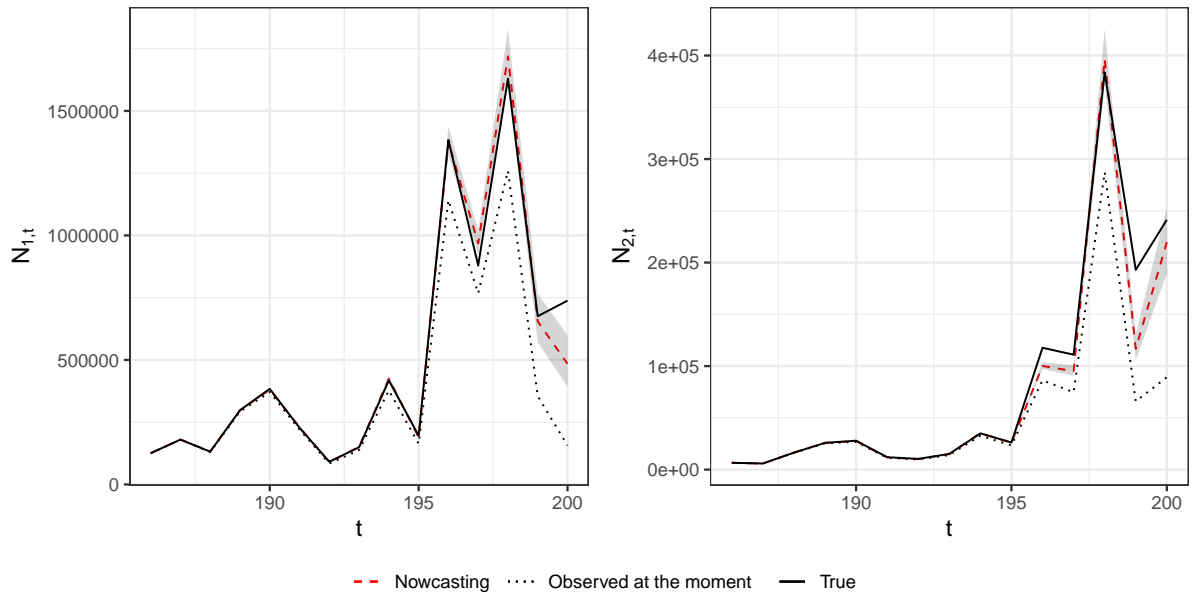


Figure A.2: Nowcasting estimates according to the Poisson model, data observed at the moment of predictions, and true observed data artificially generated in the simulation study.

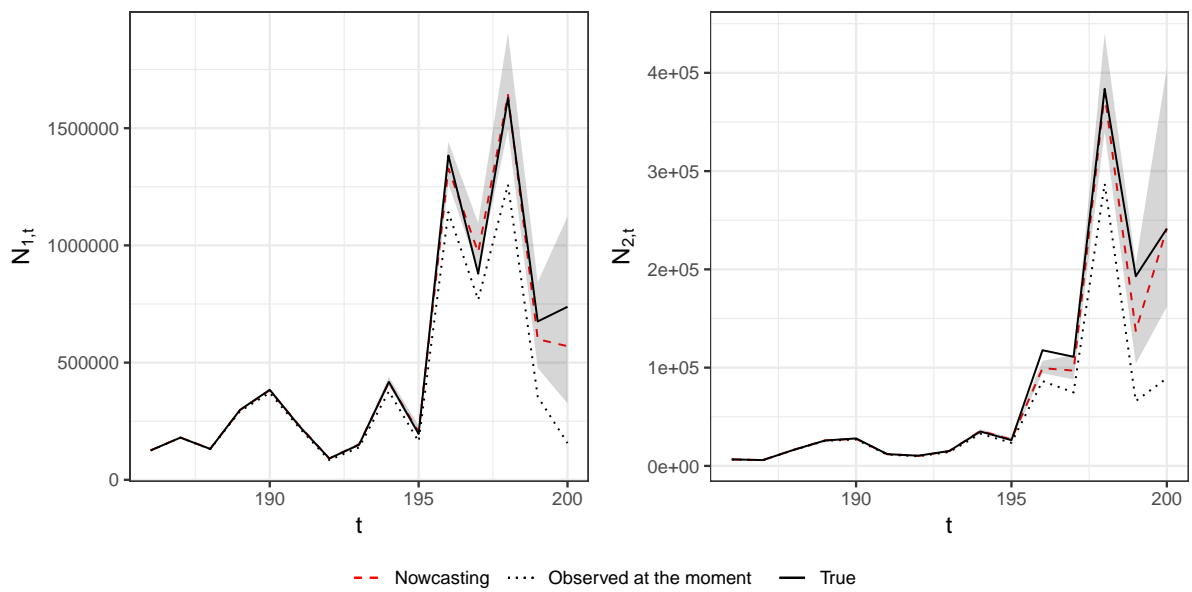


Figure A.3: Nowcasting estimates according to the independent model, data observed at the moment of predictions, and true observed data artificially generated in the simulation study.

# Appendix B

## Supplementary material for the sliding windows study

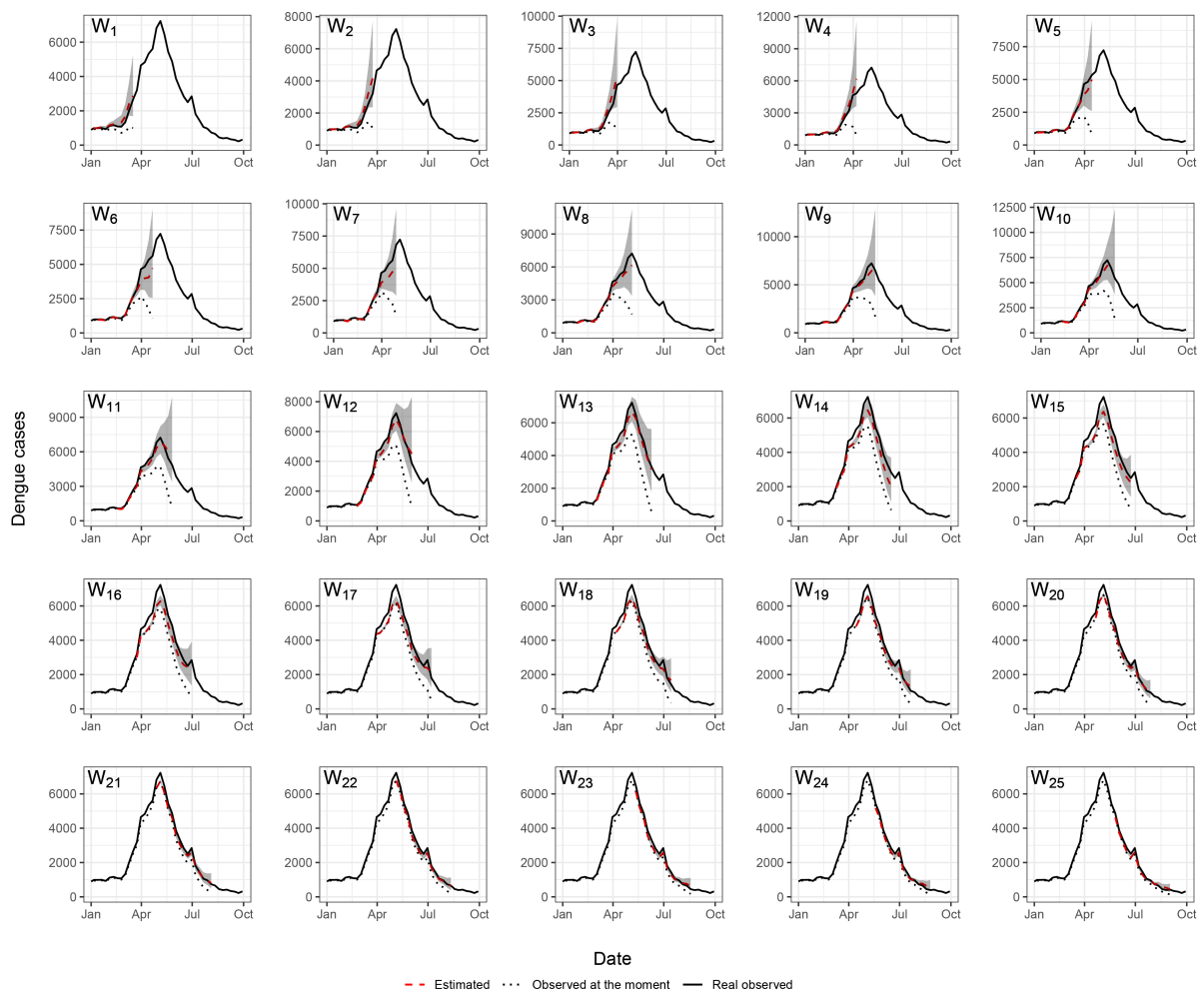


Figure B.1: Nowcasting of Dengue cases for each of the 25 windows of 70 weeks

# Appendix C

## Comparison of univariate approach and proxy approach

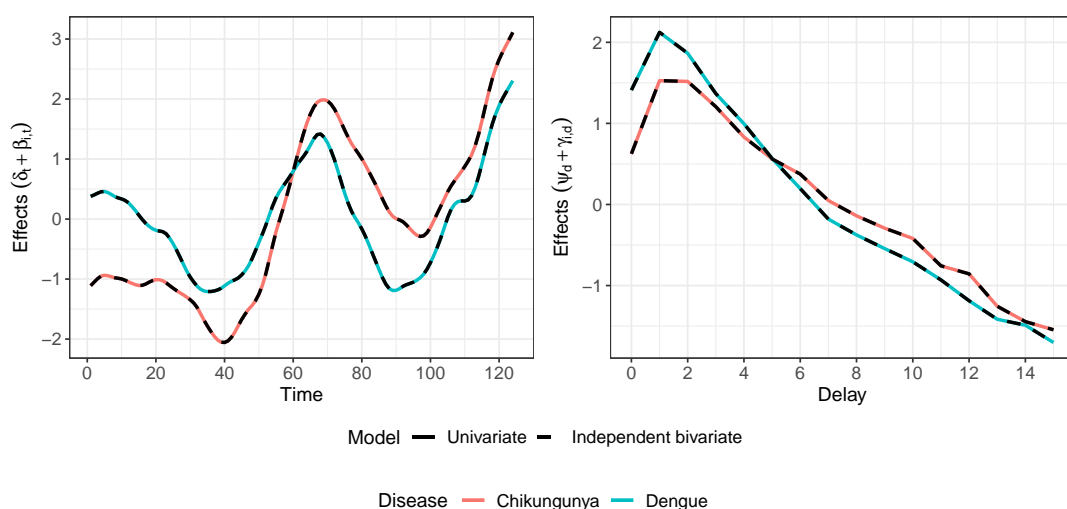


Figure C.1: Effects estimates by model.

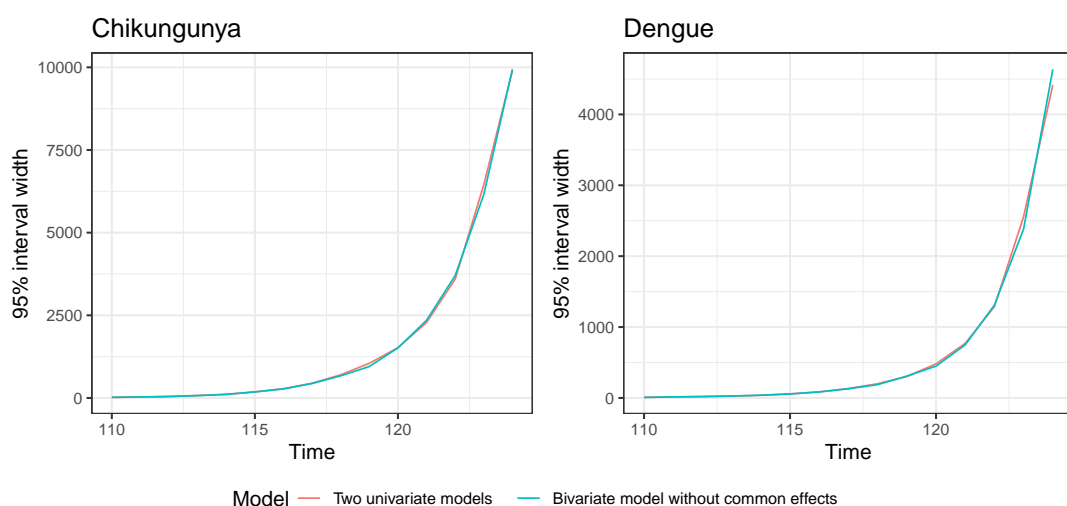


Figure C.2: 95% credible interval width according to model.

# Appendix D

## Supplementary material for the prior sensitivity analysis

Prior	Hyperparameter		
	$\phi_1$	$\phi_2$	$\sigma_\beta^{-2}$
Gamma(0.001, 0.001)	3.66 (3.35, 3.99)	4.43 (4.04, 4.87)	2848.90 (1180.54, 5824.83)
Gamma(1, $5 \times 10^{-5}$ )	3.65 (3.35, 3.98)	4.43 (4.03, 4.86)	6649.29 (2748.04, 15557.39)
Half-Normal	3.65 (3.35, 3.98)	4.43 (4.03, 4.87)	5186.75 (1898.54, 12751.65)
Half-Cauchy	3.65 (3.35, 3.98)	4.43 (4.03, 4.87)	5617.08 (1921.72, 13184.70)
Half-t	3.65 (3.35, 3.99)	4.43 (4.03, 4.87)	5378.30 (1836.43, 12788.73)
Uniform	3.66 (3.35, 3.98)	4.43 (4.03, 4.87)	5249.10 (1810.33, 12543.60)
Prior	$\sigma_\gamma^{-2}$	$\sigma_\delta^{-2}$	$\sigma_\psi^{-2}$
Gamma(0.001, 0.001)	104.59 (43.40, 256.19)	287.10 (151.71, 532.99)	9.56 (3.96, 18.94)
Gamma(1, $5 \times 10^{-5}$ )	139.05 (54.67, 302.17)	317.08 (172.17, 577.82)	10.56 (4.75, 20.19)
Half-Normal	101.60 (40.48, 238.97)	280.49 (144.78, 512.61)	8.32 (3.61, 17.56)
Half-Cauchy	105.55 (42.30, 249.81)	270.61 (147.46, 532.07)	8.38 (3.63, 17.53)
Half-t	101.57 (39.35, 238.20)	270.05 (147.01, 523.74)	8.74 (3.83, 17.80)
Uniform	103.58 (39.68, 239.21)	271.79 (145.66, 518.59)	8.41 (3.57, 17.30)

Table D.1: Hyperparameter estimates according to prior of choice.