RESEARCH ARTICLE

WILEY Statistics in Medicine

A modelling approach for correcting reporting delays in disease surveillance data

Leonardo S Bastos[1] | Theodoros Economou[2] | Marcelo F C Gomes[1] | Daniel A M Villela[1] | Flavio C Coelho[3] | Oswaldo G Cruz[1] | Oliver Stoner[2] | Trevor Bailey[2] | Claudia T Codeço[1]

[1]Scientific Computing Program, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

[2]Department of Mathematics, University of Exeter, Exeter, UK

[3]School of Applied Mathematics, Getulio Vargas Foundation, Rio de Janeiro, Brazil

Correspondence
Leonardo S Bastos, Scientific Computing Program, Oswaldo Cruz Foundation, 21040-360 Rio de Janeiro, Brazil.
Email: leonardo.bastos@fiocruz.br

Funding information
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Grant/Award Number: 88881.068124/2014-01

One difficulty for real-time tracking of epidemics is related to reporting delay. The reporting delay may be due to laboratory confirmation, logistical problems, infrastructure difficulties, and so on. The ability to correct the available information as quickly as possible is crucial, in terms of decision making such as issuing warnings to the public and local authorities. A Bayesian hierarchical modelling approach is proposed as a flexible way of correcting the reporting delays and to quantify the associated uncertainty. Implementation of the model is fast due to the use of the integrated nested Laplace approximation. The approach is illustrated on dengue fever incidence data in Rio de Janeiro, and severe acute respiratory infection data in the state of Paraná, Brazil.

KEYWORDS
Bayesian hierarchical model, dengue, INLA, reporting delay, SARI

1 | INTRODUCTION

Surveillance systems play a crucial role in managing infectious disease risk. The main requirements for a good surveillance system are timeliness, sensitivity, and specificity, together with readily interpretable outputs.[1] Timeliness reflects the speed or delay between steps in a surveillance system[2]: the time between the onset of an adverse health event and its report, and the time between report and the identification of trends or outbreaks, for example.

Disease surveillance in most countries is passive, relying on the cases reported by health care providers from patients seeking care. The number of cases reported quite commonly suffers a reporting delay that can vary across localities, being susceptible to the adherence of local health care providers to the reporting protocol, as well as the access of patients to health care. Timeliness is also affected by conflicting factors due to the disease incidence: Delays may decrease during the high-transmission season because of awareness among doctors and patients; conversely, delays may increase during high-transmission seasons because of the saturation of the health care system. Reporting delays, especially the ones whose structure varies in time, distort the relationship between the reported disease incidence and the true disease incidence. Surveillance and warning systems relying on reported incidence to assess risk can therefore be misinformed, if this delay is not somehow corrected.

From a statistical perspective, reporting delay is a censoring problem, albeit one for which the observable (reported) data will eventually become available. Note that we make a distinction here between observable data and the truth. We term observable those incidence cases that were detected and eventually reported. In disease surveillance however, data are always potentially underreported, ie, disease cases that were never detected or that were detected but never reported.

As such, the true disease count is the observable count plus any cases that were never reported. In this paper, we focus on correcting reporting delay in the observable data, noting that correcting for underreporting is generally a nontrivial task, requiring additional sources of information such as prior knowledge on underreporting rates or a sample of fully observed data (ie, the truth), as discussed in the work of Stoner et al.[3]

The aim of this paper is to propose a flexible statistical modelling framework that enables the estimation of the missing (observable) data to perform nowcasting, as well as the potential for forecasting. The framework was developed with two goals in mind: to be a useful decision making tool while, at the same time, being flexible enough to apply to a range of problems with complex data structures and to provide reliable corrections and full quantification of uncertainty. To achieve these goals, the framework should possess the following attributes.

- Practical (computational) feasibility. This is vital if the model is to be used in conjunction with a warning system that can be potentially updated in real time.
- Flexibility. The model should readily allow for covariates relating to the delay mechanism, the variability of the disease, and as other relevant information (such as Twitter feeds and weather nowcasts/forecasts).
- Complexity. The model should be able to capture any (residual) spatiotemporal variability, both in the delay mechanism and the progression of the disease. Temporal dependence is particularly important in being able to detect outbreaks.

Furthermore, with this being a prediction problem, a Bayesian formulation is desirable as it enables the use of predictive distributions that quantify all the associated uncertainty in correcting the missing values. The motivation behind such a modelling framework is presented in the model application section of this paper, where the overall goal is to develop a real-time online warning system for infectious diseases in Brazil. The model is used to correct the number of dengue cases in Rio de Janeiro and the number of severe acute respiratory infection (SARI) cases in the state of Paraná, Brazil.

The paper is structured as follows. In Section 2, we present the formulation of the problem, set the relevant notation, and discuss current approaches to model reporting delay. In Section 3, we present the modelling framework and how to perform inference to obtain the predictive distribution of the reporting cases. In Section 4, we apply the model to dengue data from Rio de Janeiro and to SARI data from the state of Paraná, Brazil. Finally, in Section 5, we provide a summary and discussion of the results obtained.

2 | BACKGROUND

2.1 | The run-off triangle

The typical data structure for the reporting delay problem is given in Figure 1, where the rows correspond to time $t = 1, 2, \ldots, T$ and the columns correspond to amount of delay (in the same units as t), $d = 0, 1, \ldots, D$, where D is the maximum possible delay. For any time step (row), the true total amount of events (eg, disease occurrences) is $N_t = \sum_{d=0}^{D} n_{t,d}$ so that $n_{t,d}$ is the number of events that occurred at time t that were reported at d time steps after t (with $n_{t,0}$ being the number that was actually reported at t). Assuming for simplicity that T is "today," then the values $n_{t,d}$ in the grey boxes of Figure 1 are missing and so are the corresponding the totals N_t. These occurred-but-not-yet-reported events are also called the run-off triangle,[4] all values of which potentially need to be estimated for accurate risk assessment (eg, for detecting a sharp increase in occurrences).

FIGURE 1 Table illustrating the typical data structure in a reporting delay problem. The values in the blue cells are fully observed number of cases at time as of time T (today), the values in grey are the occurred-but-not-yet-reported number of events (run-off triangle), and the values in red are the future number of event we may be interested to forecast [Colour figure can be viewed at wileyonlinelibrary.com]

Time	0	1	2	⋯	D-2	D-1	D	N	
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$		$n_{1,D-2}$	$n_{1,D-1}$	$n_{1,D}$	N_1	Observations
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$	⋯	$n_{2,D-2}$	$n_{2,D-1}$	$n_{2,D}$	N_2	
3	$n_{3,0}$	$n_{3,1}$	$n_{3,2}$		$n_{3,D-2}$	$n_{3,D-1}$	$n_{3,D}$	N_3	
⋮						⋮			
T-D	$n_{T-D,0}$	$n_{T-D,1}$	$n_{T-D,2}$		$n_{T-D,D-2}$	$n_{T-D,D-1}$	$n_{T-D,D}$	N_{T-D}	
T-D+1	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$n_{T-D+1,2}$		$n_{T-D+1,D-2}$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$	N_{T-D+1}	Nowcasting
T-D+2	$n_{T-D+2,0}$	$n_{T-D+2,1}$	$n_{T-D+2,2}$		$n_{T-D+2,D-2}$	$n_{T-D+2,D-1}$	$n_{T-D+2,D}$	N_{T-D+2}	
T-2	$n_{T-2,0}$	$n_{T-2,1}$	$n_{T-2,2}$		$n_{T-2,D-2}$	$n_{T-2,D-1}$	$n_{T-2,D}$	N_{T-2}	
T-1	$n_{T-1,0}$	$n_{T-1,1}$	$n_{T-1,2}$		$n_{T-1,D-2}$	$n_{T-1,D-1}$	$n_{T-1,D}$	N_{T-1}	
T	$n_{T,0}$	$n_{T,1}$	$n_{T,2}$	⋯	$n_{T,D-2}$	$n_{T,D-1}$	$n_{T,D}$	N_T	
T+1	$n_{T+1,0}$	$n_{T+1,1}$	$n_{T+1,2}$		$n_{T+1,D-2}$	$n_{T+1,D-1}$	$n_{T+1,D}$	N_{T+1}	Forecasting
T+2	$n_{T+2,0}$	$n_{T+2,1}$	$n_{T+2,2}$		$n_{T+2,D-2}$	$n_{T+2,D-1}$	$n_{T+2,D}$	N_{T+2}	
⋮				⋯					
T+K	$n_{T+K,0}$	$n_{T+K,1}$	$n_{T+K,2}$		$n_{T+K,D-2}$	$n_{T+K,D-1}$	$n_{T+K,D}$	N_{T+K}	

For reliable risk assessment at time point T, the counts in the run-off triangle need to be estimated (nowcast), ideally along with the uncertainty associated with doing so. The following section discusses some recent approaches to this problem, along with the motivation for the one proposed in this paper.

2.2 | Current approaches

The problem of reporting delay is not unique to epidemiological data. It has also been identified in actuarial science where there may be delay between insured damage and the associated insurance claim so that the challenge is to estimate the number of outstanding claims.[5] Broadly speaking, two modelling frameworks have been developed to tackle the reporting delay problem.

The first approach is to consider the distribution of the counts $n_{t,d}$ conditional on the totals N_t. The framework is then hierarchical where the N_t are assumed to be distributed as Poisson or Negative Binomial, and then, $n_{t,k}|N_t$ is multinomial with some probability vector of size D that needs to be estimated. This framework was used in a Bayesian nowcasting model to correct delays in the reporting of Shiga toxin-producing Escherichia coli in Germany.[6] The model allows for smooth changes in the temporal variation of the total number of cases N_t, as well as in the delay mechanism by characterising the multinomial probability vector as a function of time. Furthermore, a test for detecting outbreaks in infectious disease on the basis of this conditional approach has been developed.[7]

The other approach, primarily utilised in correcting insurance claims, is to think about the distribution of the cell counts $n_{t,d}$ directly. The so-called chain-ladder technique was developed as a distribution-free method to estimate the missing delayed counts.[4] Later, it was shown that the underlying model for the chain-ladder technique is a generalised linear model for $n_{t,d}$, where the mean is characterised as $\mathbb{E}[n_{t,d}] = \lambda_{t,d} = \mu + \alpha_t + \beta_d$.[5] The model has been extended in many ways to accommodate for various parametric and nonparametric functional forms, as well as potential covariates in $\lambda_{t,d}$; see for instance the works of England and Verrall[8] and Barbosa and Struchiner.[9] It is interesting to note that the chain ladder framework can be motivated from the conditional multinomial approach, as was shown in the work of Salmon et al.[10] Assume first that the total counts N_t arise from a negative binomial distribution with some mean λ_t and dispersion parameter ϕ. This is a common assumption when modelling disease count data, where the negative binomial extends the Poisson to allow for overdispersion in data where the amount of susceptible population is not actually known, which is a common problem in observational surveillance data.[11] Second, assuming the counts in each row of Table 1 are conditionally multinomial, $\boldsymbol{n}_t \sim MN(\boldsymbol{\pi}_t, N_t)$; then, it can be shown that the marginal distribution of each $n_{t,d}$ is a negative binomial with mean $\pi_{t,d}\lambda_t$ and dispersion parameter ϕ. In this way, the chain ladder method, which directly models the marginals as negative binomial, can be justified from the conditional multinomial approach (noting however that $\pi_{t,d}$ and λ_t cannot be separated).

Here, we extend the chain ladder approach with negative binomial marginals to allow for spatiotemporal variation in the counts, as well as covariate effects. Spatial variation is something that has not yet been considered in the various approaches to date; however, it is important to appreciate that both the delay mechanism and the temporal variability in the process giving rise to the counts can vary in space. For instance, in the application of the model in Section 4.2, the delay mechanism in reporting of SARI in Brazil is allowed to vary in space to account for the differences in the reporting process across administrative regions and to also borrow information across these regions. Furthermore, the particular formulation of the model that we propose readily allows for dependence along both the columns and rows of Figure 1 to capture the temporal variability of the disease occurrence and the temporal structure of the delay mechanism.

Delay	0	1	2	3	4	5	6	7	8	9	10
0	0.110	0.200	0.164	0.097	0.099	0.048	0.129	0.169	0.263	0.660	0.660
1		0.164	0.302	0.103	0.131	0.104	0.152	0.091	0.105	0.296	0.630
2			0.146	0.251	0.133	0.128	0.114	0.057	0.044	0.256	0.670
3				0.212	0.267	0.043	0.036	0.036	0.022	0.059	0.440
4					0.223	0.130	0.117	0.050	0.063	0.020	0.432
5						0.138	0.170	0.025	0.028	0.038	0.199
6							0.166	0.114	0.061	0.169	0.260
7								0.092	0.110	0.238	0.192
8									0.122	0.430	0.068
9										0.673	0.660
10											0.811

TABLE 1 Lower tail area probabilities quantifying how well the model captured the sample covariance of each column in the data. Only the upper triangular is shown, as the matrix is symmetric

Note that all the approaches mentioned here are purely statistical, in the sense that there is no specific component in the models relating to the disease dynamics. Incorporating mechanistic or physical elements relating to the disease can greatly improve predictions by allowing the science to effectively inform the modelling. Approaches combining mechanistic models such as SIR (Susceptible-Infectious-Recovered) and statistical ones have been utilized in the past to model disease time series (eg, the works of Finkenstädt and Grenfell[12]); however, such modelling efforts require well-documented data and can be computationally expensive. As mentioned, the focus here is on observational surveillance studies (often lacking in vital information such as the amount of susceptible population) and efficient modelling for use in real-time decision making.

3 | MODEL SPECIFICATION

Recall that $n_{t,d}$ is a random variable describing the number of events that occurred at time $t = 1, 2, \ldots, T$ but not reported until $d = 0, 1, 2, \ldots, D$ time units later. T is the last time step for which data is available, and D is the maximum acceptable delay, which for disease applications is potentially infinite, but for simplicity, we assume that D is bounded (this could also be true for insurance claims that must be filed within a certain period of the event). We model $n_{t,d}$ with a (conditional) negative binomial distribution with mean $\lambda_{t,d}$ and scale parameter ϕ, ie,

$$n_{t,d} \sim \text{NegBin}(\lambda_{t,d}, \phi), \qquad \lambda_{t,d} > 0, \quad \phi > 0. \tag{1}$$

The parameterisation used here is such that $\mathbb{E}[n_{t,d}] = \lambda_{t,d}$ and $\mathbb{V}[n_{t,d}] = \lambda_{t,d}(1 + \lambda_{t,d}/\phi)$. As mentioned in Section 1, we take a Bayesian approach so that predictive distributions of $n_{t,d}$ for any t and d (given the data) are readily available, as well as all the associated uncertainty in their estimation. As the dispersion parameter ϕ approaches infinity, the negative binomial reduces to the Poisson distribution. As such, ϕ can be thought of as a parameter that adds variability and, thus, flexibility to the Poisson. We assume en exponential Exp(0.1) prior distribution for ϕ with mean 10 and standard deviation 10. This is a weakly informative prior that places more probability over smaller values of ϕ and thus assumes the preference of the negative binomial to the Poisson. The term "weakly informative prior" is used here to emphasise that the prior is not an elicited informative prior, nor a completely vague "infinite" variance prior.

To capture structured temporal variability in $n_{t,d}$, the logarithm of their mean, $\lambda_{t,d}$, is characterised as follows:

$$\log(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)} + \boldsymbol{X}'_{t,d}\boldsymbol{\delta}, \tag{2}$$

where μ is the overall mean count at the log-scale and $\boldsymbol{X}'_{t,d}$ is a matrix of temporal and delay-related covariates with associated vector of parameters $\boldsymbol{\delta}$. The random effects α_t capture the mean temporal evolution of the count-generating process, whereas the β_d capture the mean structure of the delay mechanism. These can be modelled using random walks, in the simplest case, first-order ones, ie,

$$\alpha_t \sim N\left(\alpha_{t-1}, \sigma_\alpha^2\right), \quad t = 2, 3, \ldots, T, \tag{3}$$

and

$$\beta_d \sim N\left(\beta_{d-1}, \sigma_\beta^2\right), \quad d = 1, 2, \ldots, D, \tag{4}$$

where half normal $\text{HN}(\tau^2)$ prior distributions are assumed for σ_α and σ_β. These are distributions on $[0, \infty)$ where parameter τ controls the variance. Thinking about α_t and β_d as unknown functions in time and delay, τ controls the "wiggliness" of these functions—the smaller it is, the less wiggly (or in some sense "smooth") the functions will be (ie, the smaller the first-order differences will be). Noting that these random effects influence the mean count at the log-scale, for β_d, we choose $\tau = 1$ while for α_t, we choose $\tau = 0.1$. These are weakly informative priors, reflecting our belief that first-order differences across the columns (delay) will be bigger than the first-order differences along the rows (time). In other words, the temporal trend is assumed less wiggly a priori than the delay structure, though this assumption can be overridden by data given sufficient evidence. Note also that adding temporal dependence through α_t is common in modelling disease counts[13] and allows for temporal variation in the process giving rise to the counts, other than what may be explained by temporal covariates $\boldsymbol{X}'_{t,d}$ such as weather patterns. In addition, it is worth mentioning that if it is thought that the count of infections has potentially much longer temporal memory, eg, if the infectious and incubation periods are longer than one time unit, then higher order random walks can be used.
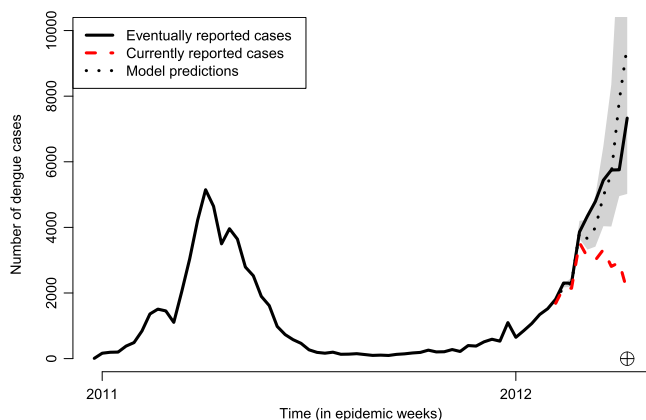
FIGURE 2 Time series of reported dengue cases in Rio de Janeiro, from January 2011 to April 2012. The black solid line shows the eventually reported number of dengue cases per week. The red dashed line shows the currently reported number of cases from the 6th to the 15th epidemic week of 2012 (circled cross). The black dotted line shows the model estimates for this period, along with 95% prediction intervals in grey [Colour figure can be viewed at wileyonlinelibrary.com]

The time-delay interaction term $\gamma_{t,d}$ is modelled as

$$\gamma_{t,d} \sim N\left(\gamma_{t-1,d}, \sigma_\gamma^2\right) \tag{5}$$

so that there is an independent realisation of a random walk order 1, for each delay column. This term is important, as it allows for changes in the delay mechanism over time. This is something we would expect because, for example, it is empirically known that delays are more severe during an outbreak due to prioritisation of treating patients rather than reporting cases. It will also indirectly allow for nonzero correlation across the columns in Figure 1, which in turn affects the variance in the totals N_t. The prior on σ_γ^2 is the same as the one on σ_α^2. Lastly, $\eta_{w(t)}$, where $w(t) = 1, \ldots, 52$ is the week index, is a seasonal component defined as a second-order random effect,

$$\eta_w \sim N\left(2\eta_{w-1} - \eta_{w-2}, \sigma_\eta^2\right), \tag{6}$$

constrained in such a way that week 1 and week 52 are joined. This term is also important as it can capture temporal variability in disease incidence that varies with the time of year. For mosquito-borne disease (such as the ones we model here), the incidence rate is strongly linked to weather variation which in turn is seasonally varying. The variance parameter σ_η^2 is less interpretable than in first-order random walks, but in general, smaller values will result in a less wiggly function. We choose a HN(1) prior to allow enough flexibility while restricting values that are too extreme. All of the components α_t, β_d, $\eta_{w(t)}$, and $\gamma_{t,d}$ are constrained to sum to zero, to allow identifiability of the intercept μ.

Figure 2 shows a time series of weekly dengue occurrences in Rio de Janeiro. This is an archetypal example of data we wish to model, exhibiting periods of very low activity but also sharp increases (outbreaks), as well as decreases. The autoregressive nature of the temporal random effects has the benefit of being able capture such behaviour in time, utilising the short-term memory in the process to adapt as new data become available. Alternative ways of characterising temporal structure such as conventional Gaussian process priors may be too smooth to capture this behaviour while models based on penalised splines can suffer from the same issue. Similarly, the autoregressive effects β_d allow for flexible characterisation of the delay mechanism (as illustrated in Section 4.) Note however that it is very important that the temporal structure is captured adequately using model checking as is performed later in Section 4. Failing that, more flexible random effect distributions can be considered, such as a mixture of Gaussian distributions.[14]

The posterior distribution for $\Theta = (\mu, \{\alpha_t\}, \{\beta_d\}, \{\gamma_{t,d}\}, \{\eta_w\}, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\eta^2, \phi)$ given all the observed data $\boldsymbol{n} = \{n_{t,d}\}$ is given by

$$p(\Theta|\mathbf{n}) \propto p(\Theta)\prod_{t=1}^{T}\prod_{d=0}^{D}p(n_{t,d}|\Theta), \tag{7}$$

where $p(n_{t,d}|\Theta)$ is the negative binomial density function (1), and $p(\Theta)$ is the joint prior distribution given by the product of the prior distributions for ϕ, σ_α^2, σ_β^2, σ_γ^2, σ_η^2, and the random effects distributions. A list of all prior distributions used, as well as a simulation experiment checking the plausibility of the data compared to simulation from the prior predictive distribution, are given as supplementary material in the following GitHub page: https://github.com/lsbastos/Delay.

3.1 | Model implementation

Samples from the posterior distribution (7) can be obtained via traditional Markov chain Monte Carlo (MCMC) methods[15] using for instance software such as NIMBLE.[16] MCMC, however, can be computationally intensive especially when the model is extended to allow for spatial variation, as discussed in the next subsection. A more efficient approach would be to obtain approximate samples from the posterior distribution of Θ using integrated nested Laplace approximation or INLA[17] using a copula approach already implemented in the INLA package for R (www.r-inla.org).

The INLA approach to obtaining samples from the posterior distribution can be significantly faster than MCMC,[18] with the added benefit of reduced user input (eg, to assess convergence of the Markov chains). Implementing the proposed model in R-INLA makes it an attractive decision making tool for correcting reporting delay in real time, eg, using an online interface for issuing warnings, as discussed in Section 4. The key concept in INLA is to combine nested Laplace approximations with numerical methods for sparse matrices for efficient implementation of latent Gaussian models. Since the model proposed here is in fact a latent Gaussian model (ie, the joint distribution of the random effects is multivariate Gaussian), it can be readily implemented using R-INLA.

3.2 | Spatial variation

In many applications, including the ones considered in this paper, the data may be spatially grouped, eg, into a number of administrative regions spanning Brazil. In general, the model presented above can be implemented independently for the various spatial regions/locations. In practice, however, it would make more sense to analyse all data together by extending the model to allow for spatial variation not only in the process giving rise to the counts but also in the delay mechanism. This allows for pooling of information to aid estimation in spatial locations with fewer data, as well as inference on how the delay mechanism varies across the different areas. The model is therefore extended to include spatial (Gaussian) random effects. Considering spatial variation where $s \in S$ denotes a spatial location or area in some spatial domain S, the model is now

$$n_{t,d,s} \sim \text{NegBin}(\lambda_{t,d,s}, \phi), \qquad \lambda_{t,d,s} > 0, \quad \phi > 0, \tag{8}$$

where $n_{t,d,s}$ is the number of occurrences in spatial location s and time point t, reported with delay d time points. In the first instance, the mean is then modelled as

$$\log(\lambda_{t,d,s}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)} + \psi_s + \beta_{d,s} + \boldsymbol{X}'_{t,d,s}\boldsymbol{\delta}, \tag{9}$$

where $\boldsymbol{X}'_{t,d,s}$ is now a model matrix that may also contain spatially varying covariates. The quantities α_t and β_d are defined in the same way as before but are now respectively interpreted as the overall temporal and delay evolution across space. The component $\beta_{d,s}$ captures the way in which the delay structure varies across space, whereas ψ_s describes the overall spatial variability and dependence in the counts. The particular formulation is motivated by the application to SARI data, where the spatial region is fairly small so the temporal effects (α_t) are a not assumed to vary with space. Given the implementation of the model in R-INLA, various possible choices exist for the specific formulation of $\beta_{d,s}$ and ψ_s. The space-time or space-delay interactions can range in complexity, from spatially and temporally unstructured Gaussian processes to nonseparable formulations (see the works of Knorr-Held[19] and Blangiardo et al[20]). The spatial effect ψ_s can be defined by an intrinsic autoregressive (IAR) process[21] if the data are counts in areal units to allow similar temporal variation in neighbouring areas. Equally, ψ_s can be defined by a stationary Gaussian process if the data are counts in point locations, eg, so that spatial dependence decreases exponentially with distance. In the application of the model in Section 4.2, where space is divided in a number of administrative areas, we use the type I space-time interaction as proposed by Knorr-Held.[19] This is a formulation where

$$\beta_{d,s} \sim N\left(\beta_{d-1,s}, \omega_\beta^2\right) \tag{10}$$

is an independent first-order random walk for each area s, and where $\psi_s = \psi_s^{\text{IAR}} + \psi_s^{\text{ind}}$, ie, the sum of a spatially structured IAR process:

$$\psi_s^{\text{IAR}} | \psi_{s' \neq s}^{\text{IAR}} \sim N\left(\frac{\sum_{s' \neq s} w_{s,s'} \psi_{s'}^{\text{IAR}}}{\sum_{s' \neq s} w_{s.s'}}, \frac{\sigma_{\text{IAR}}^2}{\sum_{s' \neq s} w_{s,s'}}\right)$$

and spatially unstructured random effects $\psi_s^{\text{ind}} \sim N(0, \sigma_{\text{ind}}^2)$. Here, σ_{IAR}^2 controls the strength of spatial dependence and σ_{ind}^2 is the variance of the spatially unstructured effects.

3.3 | Nowcasting

In any given time step T, there are a number of occurred-but-not-yet-reported (missing) values $n_{t,d}$, $t = T - D + 1, \ldots, T$; $d = 1, \ldots, D$ (grey cells in Table 1), as well as the marginal totals $N_{T-D+1}, \ldots, N_T$. Of primary interest is of course N_T, which needs to be nowcast; however, hindcasts of $N_{T-D+1}, \ldots, N_{T-1}$ may also be of interest, especially if one wants to quantify the rate of increase or decrease in the counts.

From a Bayesian perspective, this is a prediction problem where all the missing $n_{t,d,s}$ can be estimated from the posterior predictive distribution

$$p(n_{t,d,s}|\boldsymbol{n}) = \int_{\Theta} p(n_{t,d}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{n})d\boldsymbol{\Theta}, \tag{11}$$

where $\boldsymbol{n}$ denotes all the data used to fit the model. This cannot be solved analytically, however with samples from the posterior $p(\boldsymbol{\Theta}|\boldsymbol{n})$ one can use Monte Carlo to approximate (11). In practice, for each sample from $p(\boldsymbol{\Theta}|\boldsymbol{n})$, we simulate a value from the negative binomial $p(n_{t,d}|\boldsymbol{\Theta})$ to obtain an approximate sample from the predictive distribution $p(n_{t,d}|\boldsymbol{n})$. Due to the autoregressive nature of the temporal and delay components, predictions are performed sequentially starting from the top-right corner of the run-off triangle, ie, $n_{T-D+1,D}$, then moving down the rows sequentially going from left to right columnwise. Once posterior predictive samples of $n_{t,d}$ are available, then it is a matter of arithmetic to obtain equivalent samples from $p(N_t)$, the marginal totals. Samples from an approximation of the joint posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{n})$ can be obtained from R-INLA using the `inla.posterior.sample()` function as also illustrated in small-area estimation (eg, the work of Vandendijck et al[22]).

Ultimately, one has to be conscious of the approximations involved in using the INLA approach at the gain of significant increase in computational speed. In the next section, we perform a comparison between the nonspatial version of the model in (1) when implemented using both MCMC and R-INLA.

4 | MODEL APPLICATION

In this section, we apply the proposed model to two situations relating to infectious disease in Brazil. The first involves correcting reporting delay for the occurrence of dengue fever in the city of Rio de Janeiro, whereas the other relates to correcting reporting delay for SARI across the Brazilian state of Paraná. Both implementations of the model are now being used as decision making tools by local and national authorities with an associated interface to online warning systems, infoDengue (https://info.dengue.mat.br) and infoGripe (http://info.gripe.fiocruz.br).

4.1 | Dengue fever in Rio de Janeiro, Brazil

Dengue fever is an infectious vector-borne disease that has been endemic in Brazil since 1986. The transmission of dengue is characterised by significant year-to-year variability, driven by the complex interactions between environmental factors (such as temperature and humidity), human factors (such as population immunity and mobility), and viral factors (circulating strains). Uncertainties in these interactions impair the ability to prepare for and allocate resources to reduce disease burden. In this context, continuous surveillance, fast analysis, and response are key for a successful control. In principle, dengue is meant to be reported within seven days of case identification. However, in practice, less than 50% of the cases are reported within one week, less than 75% within four weeks, and no more than 90% within 9 weeks. Therefore, a reasonable upper bound for the delay time D is 10 weeks.

Reported suspected cases of dengue, as recorded by the Brazilian Information System for Notifiable Diseases (SINAN and DENGON) were provided by the Rio de Janeiro Health Secretariat. Records include two dates: date of reporting (when doctor fills in the reporting sheet) and date of digitisation (when the sheet is fed into the system), and the former is used as a reference for t. Date of disease onset, although available, presented a large percentage of missing values. The importance of timeliness for decision making is easily observed in a time series plot of dengue cases in Rio de Janeiro presented in Figure 2, where the "eventually reported" number of cases during the 2012 outbreak (black line) is considerably larger than the "currently reported" number of cases reported (red line) as of the 15th epidemic week of 2012 (from the 7th to the 13th of April 2012). Notice that if only the current number of cases is considered, a public health decision maker could potentially take wrong actions on the basis that the number of dengue cases appears to be decreasing. Also worth noting is the fact that eventually reported cases are corrected for both delays and occasionally for misclassification using laboratory confirmation tests.

4.1.1 | Results and model checking

The available data consist of weekly counts of the number of dengue cases in Rio for the time period January 2011 to December 2012, along with the associated delay information. The model used for estimation is the one given by (1) with $D = 10$ and $X_{t,d} = \mathbf{0}$ as no covariate information was available. A single run of the model was performed with time T being the 15th epidemic week of 2012 (see Figure 2), meaning that $t = 1, \ldots, 68$ weeks and $T = 68$. The model was used to correct the total number of cases N_T in that particular week, but also for the 9 weeks preceding it, as shown in Figure 2 (black dotted line), along with 95% prediction intervals. The plot indicates that the predictions actually identify the fact that there is an outbreak.

To ensure the model provides a good fit to the data, we conduct a series of checks. First, the predictive distributions of the totals $N_1, \ldots, N_{68}$ are computed from summing the respective $n_{t,d}$ over d. Figure 3 shows the predicted N_t defined as the means of these distributions, plotted against the observed N_t sorted in ascending order. The 95% prediction intervals are also added, and the plot indicates that the model estimates capture the rank of the observed values very well, bearing in mind that 10 of the 68 values are based on data the model has not seen.

Furthermore, we look at the sample mean, median, and variance of the totals N_t and check how well these are captured by computing the respective posterior predictive distributions of these three statistics (from predictive samples of the totals). The plots in Figure 4 indicate that the three statistics are well captured (ie, are not extreme with respect to the distributions). In addition, to check whether temporal dependence in N_t is well captured, we consider the sample auto-correlation in the N_t for the first eight lags. Figure 5 shows that these are well captured by the model because none of the observed values (vertical lines) is extreme with respect to the respective predictive distributions.

A further aspect of the data we would like to ensure the model captures well is the covariance $\text{Cov}(n_{t,d}, n_{t,d'})$ of the various columns in the data matrix (Figure 1). To that end, we produce many replicates of the data matrix from the respective predictive distributions of $n_{t,d}$, from which we compute the predictive distribution of the sample covariance between each column. We then compute the lower-tail area probability of the observed sample covariance value (as in Figure 4). Table 1 shows these tail area probabilities, noting that only 3% of these are extreme, ie, smaller than 0.025 or larger than 0.975, indicating that the covariances are well captured.
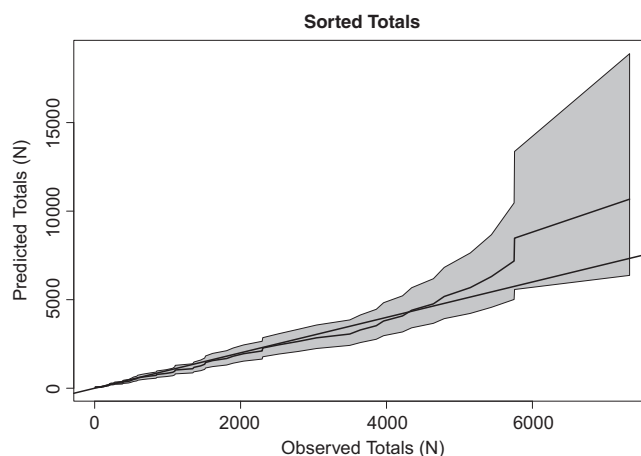

FIGURE 3 Predicted totals plotted against the respective observed (sorted) values
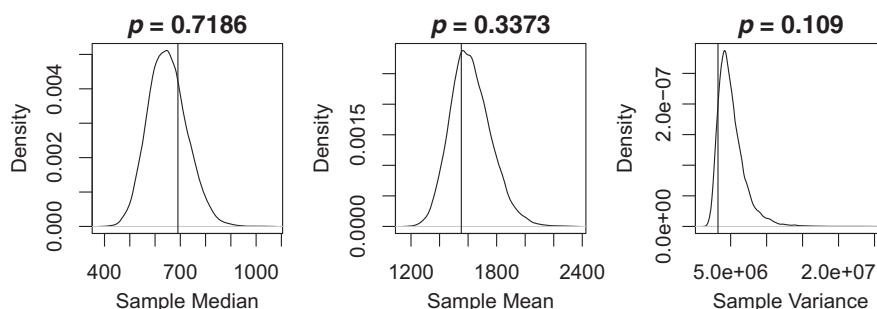
FIGURE 4 Predictive distributions for the sample mean, median, and variance of the totals N_t. Vertical lines indicate the observed values, whereas the quoted probabilities indicate the tail area of the observed values (values less than 0.025 or over 0.975 indicate that the observed value is not well represented by the model)
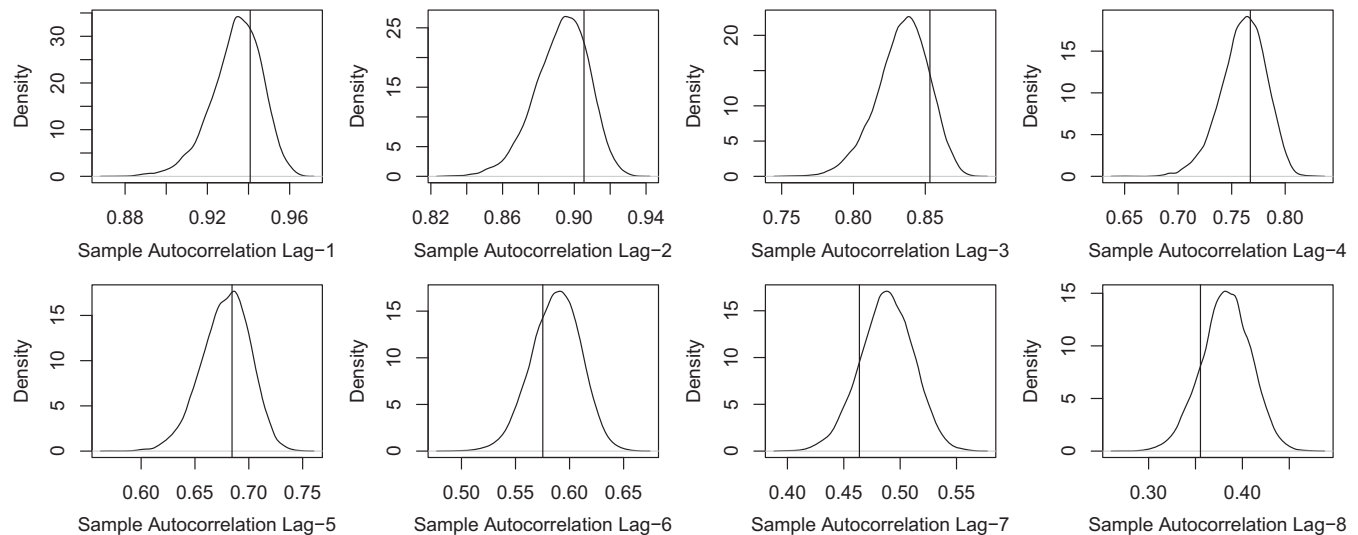
FIGURE 5 Predictive distributions for the sample autocorrelation of the totals N_t for the first eight lags. The vertical lines indicate the observed values

### 4.1.2 | Sensitivity analysis of using INLA for prediction

As discussed in Section 3.3, predictions from R-INLA are approximate and it therefore makes sense to assess the effect of the approximations. For the same data depicted in Figure 2, the model used in the previous subsection is implemented using MCMC. More specifically, the R package NIMBLE[16] was used, which uses a combination of Gibbs sampling and the Metropolis-Hastings algorithm. Three chains were run for 2.5 million iterations each, a burn-in of 2 million and thinning of 10 (totalling 150 000 samples), to ensure convergence and good mixing. Convergence was assessed by visual inspection of trace plots, and by computing the potential scale reduction factor (PSRF[23]). This compares the variance between the MCMC chains to the variance within the chains. A PSRF of 1 is obtained when the two variances are the same, so starting the chains from different initial values and obtaining a PSRF close to 1 (typically taken to be less than 1.05) give a good indication of convergence to the posterior distribution. Out of the 884 random effects and hyperparameters in the model, the maximum PSRF was 1.034, indicating all have converged according to this measure. In addition, we compute the effective number of samples per hyperparameter, by accounting for the autocorrelation in the chains. The smallest effective number of samples was 1420, ensuring that there are enough independent samples for conducting inference based on Monte Carlo.

The R-INLA and MCMC samples from the predictive distributions of the 10 unknown total counts N_t are compared using Q-Q plots shown in Figure 6. To also compare some of the individual counts $n_{t,d}$, the last two panels on the bottom right of Figure 6 compare samples from the predictive distributions of $n_{T,D-1}$ and $n_{T,D}$, ie, the last two entries of the "T" row in Table 1. Overall, the predictive distributions match well with the exception that R-INLA samples sometimes tend to slightly underestimate the upper tail. This conclusion is representative of further comparisons across other weeks not shown for conciseness. We feel that this is a reasonable compromise to the gain in computational speed with the R-INLA model taking a matter of seconds compared to hours of MCMC (per chain).

4.1.3 | Estimates and rolling predictions

Figure 7 shows estimates of three components, namely, α_t the overall temporal evolution in the counts, β_d the delay structure, and $\eta_{w(t)}$ the seasonal variability. The overall temporal effect is increasing at first but then plateaus, perhaps reflecting an increase in the susceptible population. The delay structure is almost linear and decreasing, as would be expected—the more time goes by, the more cases are being reported. There is also a strong seasonal component capturing an increase in the early part of the year and a decrease later on.

Furthermore, Figure 8 shows weekly rolling predictions, starting from the 15th going to the 26th epidemic week of 2012. This period was chosen specifically to test the ability of the model to capture the outbreak as well as the relatively sharp decline, in the eventually reported number of cases (black line). From Figure 8, it is evident that the model (black dotted line) captures both the increase and decrease in the eventually reported number of cases, despite a sharp decrease in the
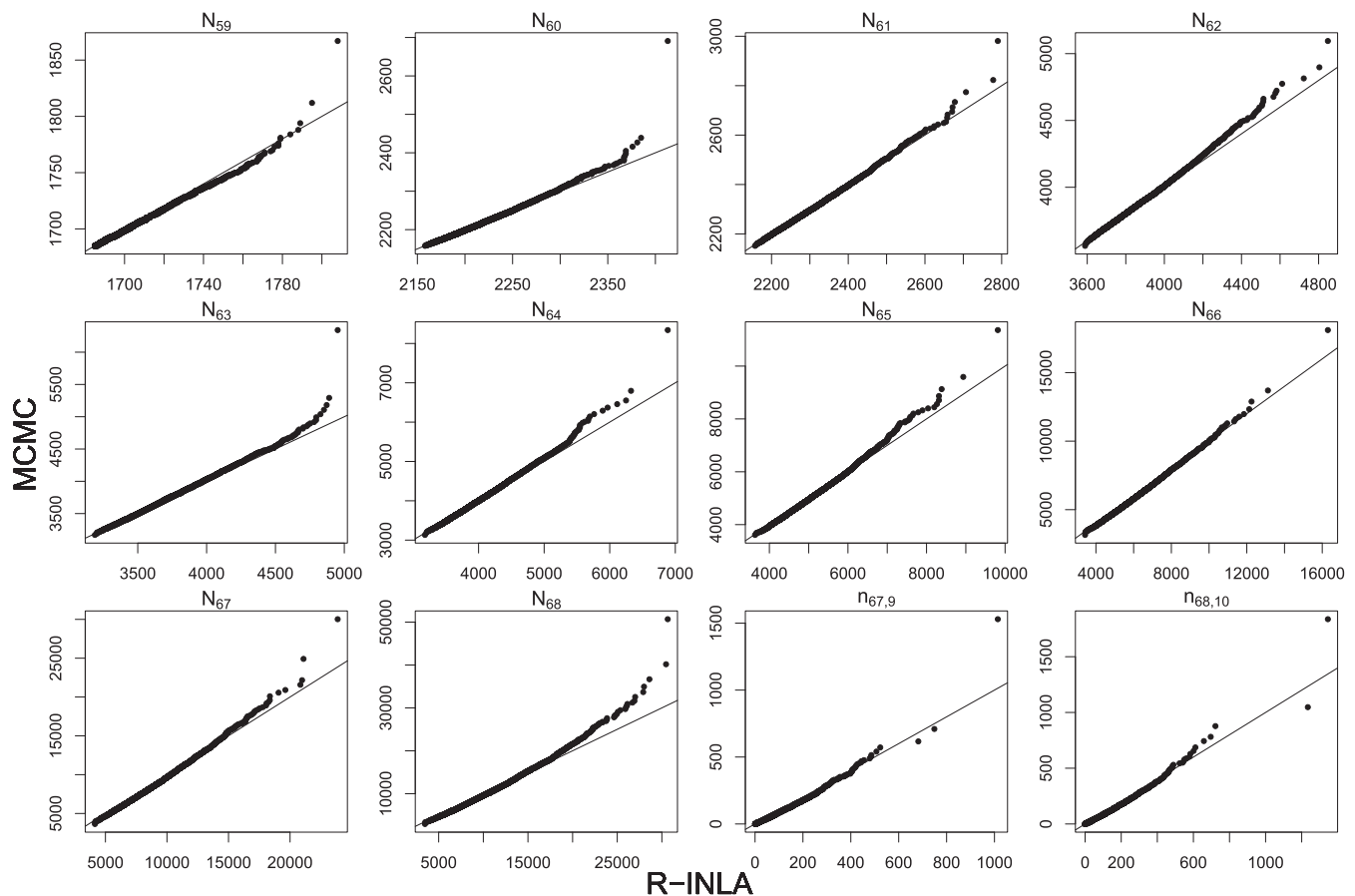
FIGURE 6 Q-Q plots comparing R-INLA and MCMC samples from the predictive distribution of the total counts N_t for $t = 59, \ldots, 68$, where $T = 68$ is the 15th epidemic week of 2012. INLA, integrated nested Laplace approximation; MCMC, Markov chain Monte Carlo
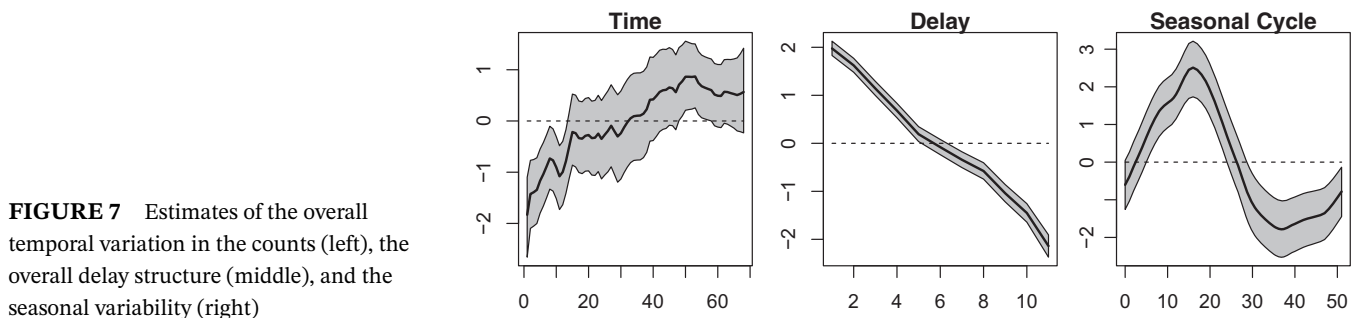

FIGURE 7 Estimates of the overall temporal variation in the counts (left), the overall delay structure (middle), and the seasonal variability (right)

week that the peak occurs (top-right panel). It is important to note that most of the eventually reported counts are within the 95% prediction intervals, particularly for time T (indicated by the circled cross), which is the most important value.

Predictions from the particular model implementation presented here are currently being used to inform a disease warning system in 790 Brazilian municipalities from six Brazilian states: Rio de Janeiro, Paraná, Espírito Santo, Ceará, Minas Gerais, and São Paulo. The warning system also uses Twitter feeds and weather information, but the primary source of information for predicting both dengue and chikungunya cases (another mosquito-borne disease) is the posterior predictive means from the model presented here. Note that the model is being fitted independently in each region, but in the next section, an application to SARI involves a spatial structure.

4.2 | SARI in Paraná, Brazil

The lack of a baseline for detecting changes in disease severity during the 2009 H1N1 Influenza pandemic led the World Health Organization (WHO) to standardise, in 2011, a definition for the notification of SARI worldwide. SARI is defined
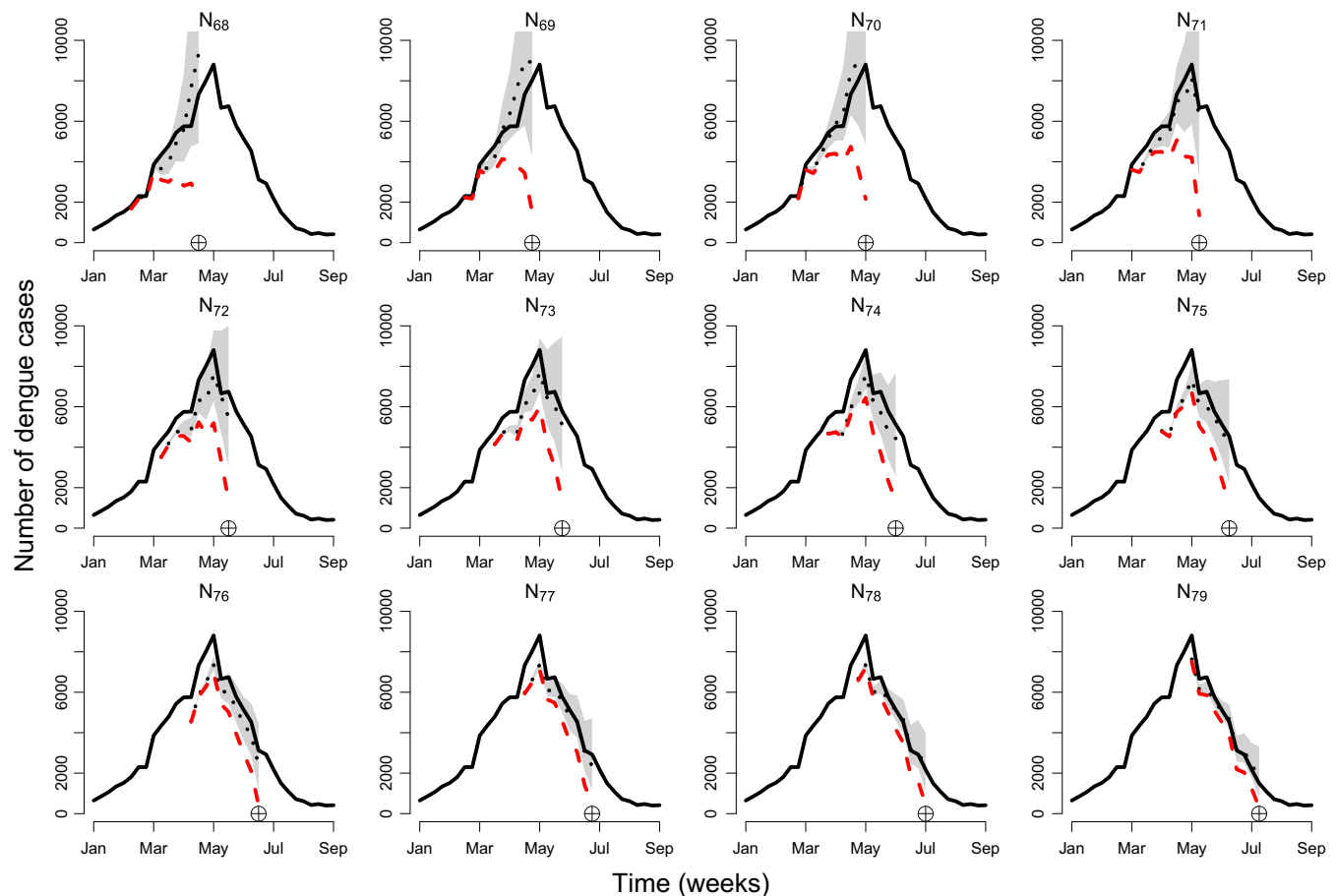
FIGURE 8 Time series of dengue cases in Rio de Janeiro for 12 epidemic weeks starting from the 15th epidemic week of 2012 on the top left ($T = 68$). The black line shows the eventually reported number of cases, the red dashed line shows the number of currently reported cases, and the black dotted line shows model predictions (of the eventually reported number of cases) along with 95% prediction intervals. The circled cross symbol indicates the epidemic week $T = 68, 69, \ldots, 79$ [Colour figure can be viewed at wileyonlinelibrary.com]

as an acute respiratory infection with onset within the past 10 days, and a history of fever or measured fever above 38°C, coughing, and requiring hospitalisation.[24] The goals of SARI surveillance include, among other things, to determine the seasonal patterns of respiratory virus circulation, to detect the emergence of high pathogenicity influenza viruses, and to provide timely information to guide prevention policies.

In general, weekly reports of SARI activity are sent from hospitals to the local (municipality) health authorities, then aggregated at the state and national levels, and eventually sent to the WHO. Besides total case counts of SARI, laboratory tests are carried out to identify the etiological agent associated and provide specific diagnosis. This procedure also enables stratification of the number of SARI cases per type of virus. Each one of those steps introduces delays in the information available for epidemiological situation rooms, created at local or global levels. For influenza, having a precise estimate of SARI activity in a timely manner is fundamental to update the indicators of activity upon which decisions are made.

In 2009, the Brazilian state of Paraná was heavily affected by the H1N1 epidemic, accounting for 52% of all reported cases in Brazil and an incidence rate of SARI at least four times greater than the other states.[25] Brazil implemented the national SARI surveillance in 2009 and since then Paraná remains among the states with largest attack rates. Paraná is an important point of entry from Argentina and Paraguay into Brazil, has an intense touristic activity, and has an important poultry industry (type of landscape at risk of emergence of new influenza viruses). As such, implementing a nowcasting SARI surveillance has strong practical implications.

4.3 | Data

The data consist of SARI reports extracted from the Brazilian Information System for Notifiable Diseases (SINAN) starting from January 1, 2016, and ending at April 2, 2017 (66 weeks), for the state of Paraná. The state is divided into

399 municipalities, and each municipality belongs to one of 22 health regions. The available data are aggregated at the health region level. The goal is to use the proposed model to correct reporting delay across the health regions, taking into account spatial variability in the delay mechanism and the disease process, as well as allow for spatial dependence in neighbouring health regions. In particular, we consider the model described in Section 3.2, namely,

$$n_{t,d,s} \sim \text{NegBin}(\lambda_{t,d,s}, \phi)$$
$$\log(\lambda_{t,d,s}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \beta_{d,s} + \psi_s^{\text{IAR}} + \psi_s^{\text{ind}} \quad (12)$$

with $t = 1, \ldots, 66$ (weeks), $d = 0, \ldots, 10$ (delay weeks), and $s = 1, \ldots, 22$ (health regions). The quantities α_t, β_d, and $\gamma_{t,d}$ are defined as before, whereas $\beta_{d,s} \sim N(\beta_{d-1,s}, \omega_\beta^2)$ allow for unstructured spatiodelay variability. Note that more complex spatiotemporal structures are possible,[13] who use a penalised splines to define space-time interaction, estimated using R-INLA. Lastly, $\psi_s^{\text{ind}} \sim N(0, \sigma_\psi^2)$ captures spatially unstructured variability while ψ_s^{IAR} [21] is spatially structured according to an IAR process with a neighbouring structure defined by a 22×22 adjacency matrix $\mathbf{W}$, where $w_{i,j} = 1$ if the health region i is an administrative neighbour of health region j, and $w_{i,j} = 0$, otherwise.

In model (12), we assume that the delay structure varies with the health region, through $\beta_{d,s}$, while the overall temporal evolution of the disease counts α_t is the same across the regions. This is because the state is fairly small and we would not expect the disease transmission to vary considerably across space. Similarly, the interaction term $\gamma_{t,d}$ is spatially constant. The term $\psi_s^{\text{IAR}} + \psi_s^{\text{ind}}$ captures overall similarity in disease counts across the health regions; however, it also allows for some regions to be different (on average) if there is such evidence in the data.

4.3.1 | Results

Figure 9 shows the weekly time series of the eventually reported SARI cases in Paraná from the first epidemic week of 2016 to the 14th epidemic week of 2017 (solid black line). The plot also shows the currently reported number of SARI cases for the last 10 weeks, up to and including the 14th epidemic week ending at April 2, 2017 (dashed red line). Finally, the plot also depicts the estimated mean of the corresponding predictive distribution from model (12), along with 95% prediction intervals (dotted black line and shaded region). The model is able to capture the increasing trend of the disease counts, and the predictions are much closer to the true value compared to the currently reported counts (which actually indicate a decline).

At epidemic week 14 during the 2017 SARI season in Brazil, the present nowcasting strategy was able to correctly detect that the SARI activity in the state of Paraná likely reached a historically defined pre-epidemic level.[26] The currently reported number of cases in weeks 13 and 14, $n_{14,0} = 29$ and $n_{13,0} + n_{13,1} = 59$, were both below the epidemic threshold of 64.6 cases (horizontal blue dashed line in Figure 9), and it was only by the end of week 16 that the currently reported number cases of week 13 went above the threshold. In other words, the model was able to detect that epidemic activity
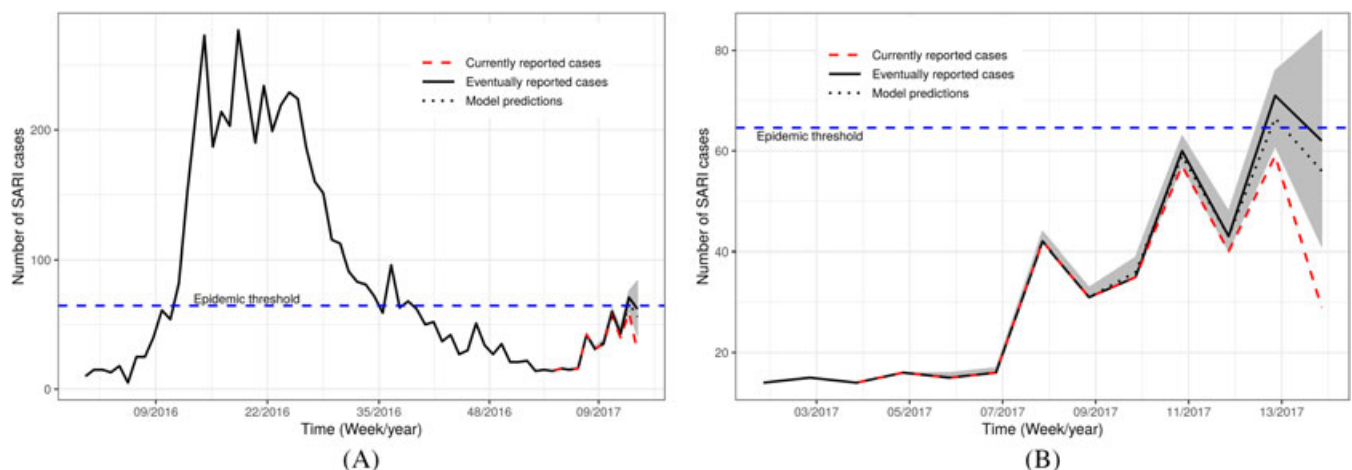
FIGURE 9 Time series of severe acute respiratory infection (SARI) cases reported in the whole of Paraná state. The black solid line shows the true number of SARI cases per week. The red dashed line shows the number of cases that were reported at the 14th epidemic week of 2017. The black dotted line shows the model estimates along with 95% prediction intervals. Subfigures (A) and (B) differ only on the time scale, where (A) starts from January 2016, whereas (B) starts from January 2017 [Colour figure can be viewed at wileyonlinelibrary.com]
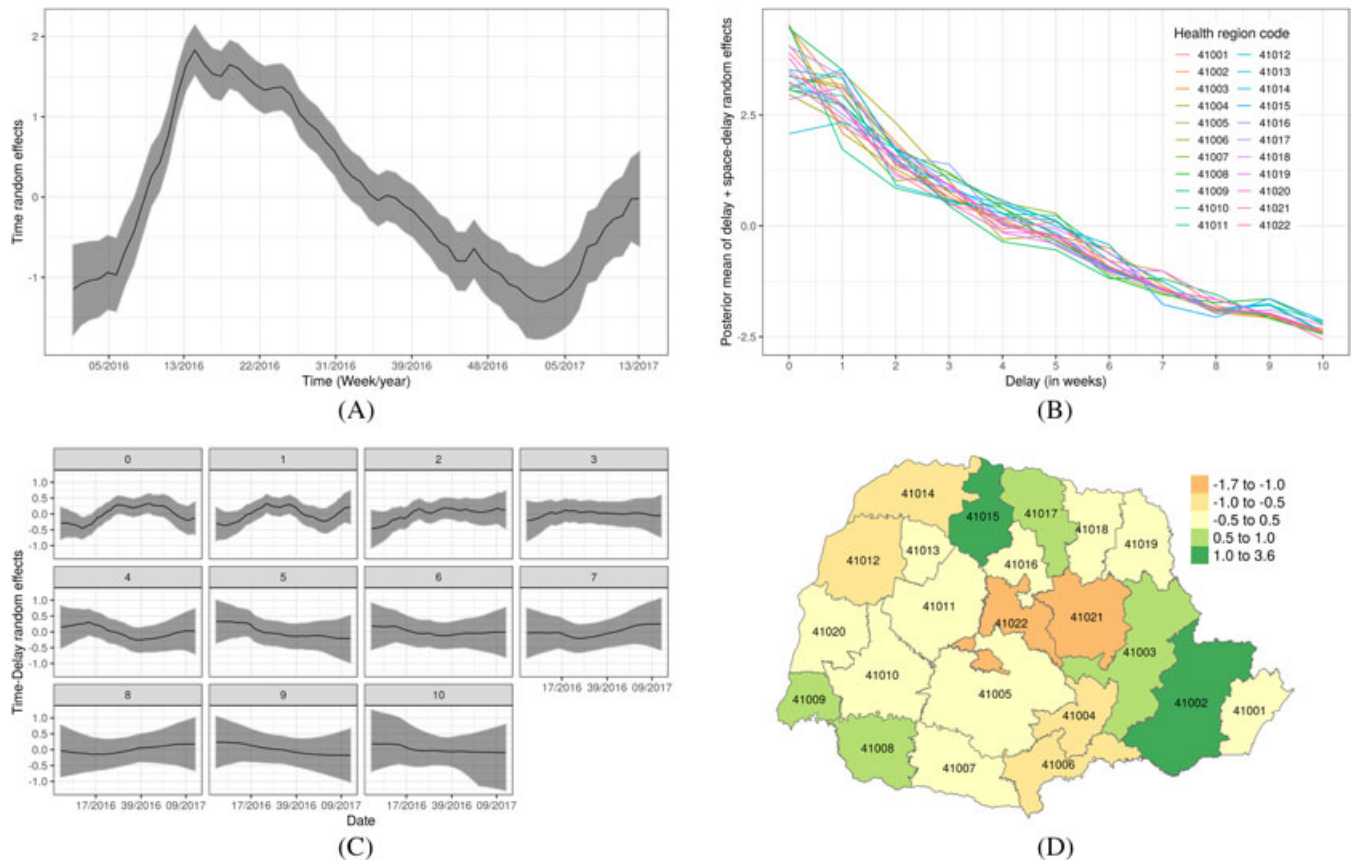
FIGURE 10 Estimates of the various random effects. A, Posterior mean with 95% credible intervals for time random effects α_t; B, Posterior mean of the space-delay random effects $\beta_d + \beta_{d,s}$ by health regions; C, Posterior mean of the time-delay random effects $\gamma_{t,d}$ by delayed weeks; D, Posterior mean of the spatial random effects ψ_s [Colour figure can be viewed at wileyonlinelibrary.com]

started effectively one week after it did, while it took three weeks for the official data to detect it. In practice, this means that our system was able to detect the qualitative transition two weeks earlier, which could have been used by public health authorities to trigger mitigation strategies at the population and health practitioner level.

Figure 10A shows the estimate of the temporal evolution α_t. Severe acute respiratory infection is very sensitive to weather variations, and the state of Paraná, which is located in the south of Brazil, has well-defined seasons with spring and winter being the seasons associated with the majority of SARI reports. This is reflected in the estimate of α_t (posterior mean and 95% credible intervals). Estimates of the delay mechanism, which is different across different health regions ($\beta_{d,s}$), are shown in Figure 10B. On average, the mean reporting count decreases with delay (in weeks); however, there is considerable variability across the health regions, particularly during the first two weeks. This reflects the fact that delay is likely related to several factors such as the region infrastructure, which varies considerably in space.

Estimates of the time-delay interaction term $\gamma_{t,d}$ are shown in Figure 10C. The plots show that the temporal evolution for $d = 0$ (no delay) and $d = 1$ (1 week delay) is negative in the first quarter of each year, suggesting that possible awareness of the SARI epidemic leading to faster notifications when a case is known. Furthermore, Figure 10D shows the estimate of the overall spatial variability term $\psi_s^{\text{IAR}} + \psi_s^{\text{ind}}$. This indicates some variability in the number of SARI reports across the regions, but also similarity in neighbouring regions. This is probably reflecting unobserved factors relating to the susceptible population (including population size). In order to assess whether spatial correlation was adequately captured, we consider the measure

$$R = \frac{\text{var}\left(\psi_s^{\text{IAR}}\right)}{\text{var}\left(\psi_s^{\text{IAR}} + \psi_s^{\text{ind}}\right)}.$$

This quantifies the contribution of the structured random effect ψ_s^{IAR} to the total variance of the spatial effect $\psi_s^{\text{IAR}} + \psi_s^{\text{ind}}$. Values close to zero indicate there is not much spatial correlation while values around 0.5 indicate roughly equal contribution for structured and unstructured spatial effects. Higher values, which can be greater than 1 due to possible nonzero correlation between ψ_s^{IAR} and ψ_s^{ind}, indicate the structured random effects are capturing most of the
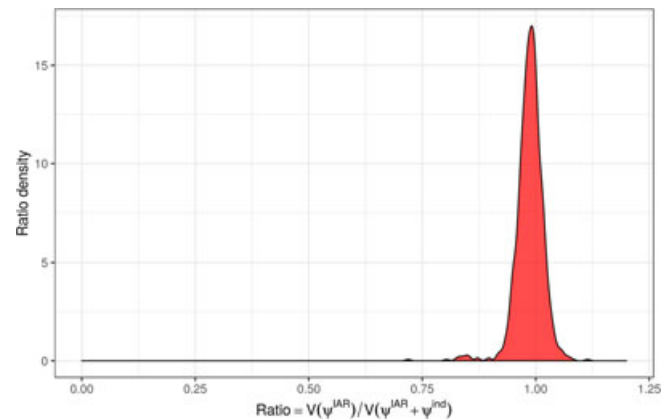
FIGURE 11 Posterior distribution of R [Colour figure can be viewed at wileyonlinelibrary.com]

variability. Figure 11 shows a plots of the posterior distribution of R, which is centred at 1, indicating that the ψ_s^{IAR} explains most of the variance with minimal contribution from ψ_s^{ind}. Had the structured random effect not be capturing spatial correlation adequately, we would expect more contribution from the unstructured effect (which can potentially compensate).

The R code and data for reproducing the dengue and SARI analyses are available on the aforementioned GitHub webpage: https://github.com/lsbastos/Delay.

5 | DISCUSSION

We have presented a general modelling framework and implementation method to flexibly model reporting delays that can in principle be applied to any disease. In fact, the proposed framework can be applied to any reporting system for which the data are described by a run-off triangle given in Figure 1. The model was illustrated using dengue data from Rio de Janeiro and SARI data in the state of Paraná in Brazil.

The two case studies, dengue and SARI, have demonstrated that the framework has desired flexibility and complexity. In the application to dengue, a model with a dependency structure in both time and delay was utilised, whereas in the case of SARI data, spatial variability and dependence was assumed in order to borrow information across the spatial units and to allow for the (arbitrary) division of the data in health regions. Although none of the models included any covariates, this is a fairly trivial task in the proposed R-INLA implementation.

The implementation of the models in the Bayesian framework is extremely fast because we make use of the Laplace approximation (INLA) to compute samples from the (marginal) posteriors. In fact, the model fitted to the dengue data is currently being used to nowcast dengue cases for use in a warning system in Brazil called Info-Dengue (https://info.dengue.mat.br/). Furthermore, nowcasts from the same model are being directly used to produce warnings for influenza and SARI across the whole of Brazil by the Ministry of Health, where, for instance, the 2017 SARI outbreak in Paraná state was anticipated 2 weeks earlier using our proposed method. Accurate estimates of the number of disease cases are of utmost importance to avoid misclassification, eg, failing to issue a high incidence alert. Therefore, this general method can greatly help warning analysts in surveillance systems to making well-informed decisions. Furthermore, the availability of samples from the predictive distribution of the counts implies that the predictions from the proposed models can be readily utilised in a decision theoretic framework for issuing warnings (as in the work of Economou et al[27]).

CONFLICT OF INTEREST

All contributing authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The dengue fever and SARI datasets, and the R code to reproduce the analyses of this paper are available at https://github.com/lsbastos/Delay.

ORCID

Leonardo S Bastos 🄳 https://orcid.org/0000-0002-1406-0122

REFERENCES

1. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. *J Royal Stat Soc, Ser (Stat Soc)*. 1996;159(3):547-563.
2. Klaucke DN, Buehler JW, Thacker SB, Parrish RG, Trowbridge FL, Berkelman RL. Guidelines for evaluating surveillance systems. *Morb Mortal Wkly Rep*. 1988;37(Suppl5):1-18.
3. Stoner O, Economou T, Marques da Silva GD. A hierarchical framework for correcting under-reporting in count data. *J Am Stat Assoc*. 2019.
4. Mack T. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*. 1993;23(2):213-225.
5. Renshaw AE, Verrall RJ. A stochastic model underlying the chain-ladder technique. *Br Actuar J*. 1998;4(04):903-923.
6. Höhle M, an der Heiden M. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*. 2014;70(4):993-1002. https://doi.org/10.1111/biom.12194
7. Noufaily A, Farrington P, Garthwaite P, Enki DG, Andrews N, Charlett A. Detection of infectious disease outbreaks from laboratory data with reporting delays. *J Am Stat Assoc*. 2016;111(514):488-499. https://doi.org/10.1080/01621459.2015.1119047
8. England P, Verrall R. Stochastic claims reserving in general insurance. *Br Actuar J*. 1993;8(3):443-518.
9. Barbosa MTS, Struchiner CJ. The estimated magnitude of AIDS in Brazil: a delay correction applied to cases with lost dates. *Cadernos de Saúde Pública*. 2002;18:279-285.
10. Salmon M, Schumacher D, Stark K, Höhle M. Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal*. 2015;57(6):1051-1067. https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201400159
11. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*. 2005;5(3):187-199.
12. Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. *J Royal Stat Soc: Ser C (Appl Stat)*. 2000;49(2):187-205. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00187
13. Bauer C, Wakefield J, Rue H, Self S, Feng Z, Wang Y. Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statist Med*. 2016;35(11):1848-1865. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6785
14. Faulkner JR, Minin VN. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*. 2018;13(1):225-252. https://doi.org/10.1214/17-BA1050
15. Gamerman D, Lopes HF. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: CRC Press; 2006.
16. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J Comput Graph Stat*. 2017;26:403-413.
17. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Royal Stat Soc: Ser B (Stat Methodol)*. 2009;71(2):319-392. https://doi.org/10.1111/j.1467-9868.2008.00700.x
18. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian computing with INLA: a review. *Annu Rev Stat Appl*. 2017;4(1):395-421.
19. Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statist Med*. 2009;19(17-18):2555-2567.
20. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal epidemiology with R-INLA. *Spatial Spatio-temporal Epidemiol*. 2013;4:33-49.
21. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*. 1991;43(1):1-20. https://doi.org/10.1007/BF00116466
22. Vandendijck Y, Faes C, Kirby RS, Lawson A, Hens N. Model-based inference for small area estimation with sampling weights. *Spatial Statistics*. 2016;18:455-473.
23. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. 1998;7(4):434-455.
24. Fitzner J, Qasmieh S, Mounts AW, et al. Revision of clinical case definitions: influenza-like illness and severe acute respiratory infection. *Bull World Health Organ*. 2018;96(2):122.
25. Codeço CT, da Silva Cordeiro J, da Silva Lima AW, et al. The epidemic wave of influenza a (H1N1) in Brazil, 2009. *Cadernos de Saúde Pública*. 2012;28(7):1325-1336.
26. Vega T, Lozano JE, Meerhoff T, et al. Influenza surveillance in Europe: establishing epidemic thresholds by the moving epidemic method. *Influenza Other Respir Viruses*. 2013;7(4):546-558.
27. Economou T, Stephenson DB, Rougier JC, Neal RA, Mylne KR. On the use of Bayesian decision theory for issuing natural hazard warnings. *Proc Royal Soc Lond A: Math Phys Eng Sci*. 2016;472(2194):20160295. http://rspa.royalsocietypublishing.org/content/472/2194/20160295