A Bayesian Hierarchical Model for COVID-19 Cases in Mindanao Philippines

Jejemae D. Nacion[1], Bernadette F. Tubo Ph.D.[2]

[1,2] College of Science and Mathematics
Department of Mathematics and Statistics
MSU-Iligan Institute of Technology, 9200 Iligan City, Philippines
jejemae.nacion@g.msuiit.edu.ph[1], bernadette.tubo@g.msuiit.edu.ph[2]

Abstract

In this study, a Bayesian hierarchical modelling approach is utilized to nowcast COVID-19 cases in Mindanao, Philippines. The proposed methodology explores the possibility of a flexible way of correcting the time and space delayed reports of the COVID-19 cases for a duration of 4 weeks for the 27 provinces in Mindanao.

Consider the model $log(\lambda_{t,d,s}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \beta_{d,s} + \psi_s^{IAR} + \psi_s^{ind}$. To estimate the parameters of the given model, a Bayesian approach is considered. Moreover, the focus of the modelling approach is to include some parameters that will correct reporting delays in the dataset and derive a model using the Integrated Nested Laplace Approximation (INLA).

Exploring and fitting the said model using the COVID-19 counts as reported by DOH in the 27 provinces of Mindanao show that the proposed model was able to capture the increasing trend of the COVID-19 disease counts. Moreover, it was observed that the prediction counts are much closer to the true count compared to the currently reported counts which actually showed a decline.

Keywords– Bayesian Hierarchical; INLA; Spatio-temporal Model; COVID-19; Reporting Delay; Nowcasting

1 Introduction

Epidemiological surveillance is the ongoing and systematic collection, analysis, and interpretation of health data in the process of describing and monitoring a health event [14]. Timeliness, which relates to the speed or delay between actions in a monitoring system, is one attribute of effective surveillance. Reporting delays are well-known issues that breach timeliness. Because of defects or ”lags” in the data collection method, the available count data, for a time, represents the truth less accurately [24]. The associated bias from delayed reports affects parameter estimates, predictions, and statistical inferences. This added uncertainty could reduce the confidence of the policymakers and warning systems in the public health decision-making process [22, 23].

In the Philippines, the currently experienced COVID-19 disease highlighted the problem of disease surveillance. COVID-19, given its complexity and behavior, exposed the problem of delayed reporting on disease occurrences. Reporting delay is affected by conflicting factors due to the disease incidence: (1) a prolonged interval between the time an individual recognizes symptoms and is able to seek care and receive confirmatory testing, (2) administrative backlogs and delays in the acquisition, processing, and ultimate reporting of information, and (3) the length of time necessary to conduct a full case investigation [15]. However, significant choices

should be made continuously notwithstanding the way that the latest data is likely incomplete. Hence, on that account, the methodology is needed to help provide a clearer picture to decision-makers in the face of the uncertainty from delays in reporting.

Studies related to reporting delays have been introduced in the past by the authors Brookmeyer and Damiano [5] and Kalbfleisch and Lawless [10], who both dealt with back-calculation and initiating events (events in the past) for AIDS incidence in 1989. Lawless in 1994 also dealt with the estimation allowing random temporal fluctuations in reporting delays [17]. In recent studies, Kassteele et al. 2019 [12], Stoner et al. 2020 [22], McGough et al. 2020 [18], and Kline et al. 2021 [15] enhance the model on its flexibility and interpretability, and extended prior works within a Bayesian framework. Also, Stoner and Economou 2019 [23] and Rotejanaprasert et al. 2020 [20] incorporate spatial dependence into temporal models using a Bayesian framework with sliding windows. As such, in the case of spatio-temporal models, the joint distribution would describe the behavior of the process at all spatial locations and at all times. In the study of Bastos et al. 2019 [2], a Bayesian hierarchical modelling approach was used to correct reporting delays and quantify the associated uncertainty in the missing values. Their approach is illustrated by dengue fever incidence data in Rio de Janeiro and severe acute respiratory infection data in the state of Paraná, Brazil.

Motivated by the works of Bastos et al. 2019 [2], this paper considers the Bayesian hierarchical approach for correcting report delays, suitable for a wide range of spatio-temporal count data, and applies it to counts of COVID-19 cases in the provinces of Mindanao.

2 Related Literature

The problem of occurred-but-not-yet-reported cases is well known from the HIV/AIDS outbreak. Different statistical approaches have been proposed in the past to handle delayed reporting. A standard reference is Lawless 1994 [17]. Moreover, some early contributions in application to the estimation of HIV/AIDS incidences were Lagakos et al. 1998 [16], who developed nonparametric methods, Kalbfleisch and Lawless 1989 [10] and Harris 1990 [8], who both considered Poisson processes, and Kalbfleisch and Lawless 1991 [11], who specified the regression models which led to simple tests and estimation of covariate effects based on right truncated data. The estimation procedure “back-calculation” by Brookmeyer and Gail 1988 [6] and Bacchetti et al. 1993 [1], applied to AIDS incidence data in the United States, refers to the reconstruction of the past history of first events (onset date) that must have occurred to give rise to the observed pattern of second event cases (date report confirmed), under the assumption of a known delay distribution. Furthermore, statistical approaches are proposed not only to epidemiological data but also in actuarial science like Renshaw 1998 [19] 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. In recent literature, like Höhle and an der Hieden 2014 [9], van de Kassteele et al 2019 [12], and Bastos et al 2019 [2], instead of back-calculation, the term ’nowcasting’ is often used for estimating the current number of events using only the available partial information reported.

These various approaches may be be broadly classified in two groups: one which models the delay counts $(n_{t,d})$ jointly but also conditionally on the total (N_t), in conjunction with a separate model for the total, e.g. Salmon et al 2015 [21], Höhle and an der Heiden 2014 [9], and Stoner and Economou 2019 [22]; and another where the delayed counts $(n_{t,d})$ are modelled marginally without explicitly modelling/using historical information on the totals, e.g. Bastos et al. 2019 [2] and McGough et al. 2020 [18].

Bastos et al. 2019 [2] and Rotejanaprasert et al. 2020 [20] applied this second approach to spatio-temporal SARI data from Brazil and to dengue fever data from Thailand [20], respec-

tively. It was a generalization of older chain-ladder approaches where they extended the model with negative binomial marginals to allow for spatio-temporal variation in the counts, as well as covariate effects. The approach is quite flexible, as it can potentially incorporate a wide variety of temporal, spatial and spatio-temporal structures.

Finally, as for the approximation process, Bastos et al 2019 [2] performed a comparison between the nonspatial version of the model when implemented using both MCMC and R-INLA. Approximations involved in using the INLA approach gained a significant increase in computational speed. Accordingly, it 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.

3 Methodologies

3.1 Data Description of Real Dataset

The available data consist of weekly counts of COVID-19 reports that were extracted from the Department of Health Data Drop starting from January 1, 2020, and ending on April 30, 2021 (70 weeks) for the whole island of Mindanao. The island is divided into 449 municipalities, and each municipality belongs to one of the 27 provinces. The available data are aggregated at the provincial level. The goal is to use the proposed model to correct reporting delays across the provincial level, taking into account spatial variability in the delay mechanism and the disease process, as well as allow for spatial dependence in neighboring regions.

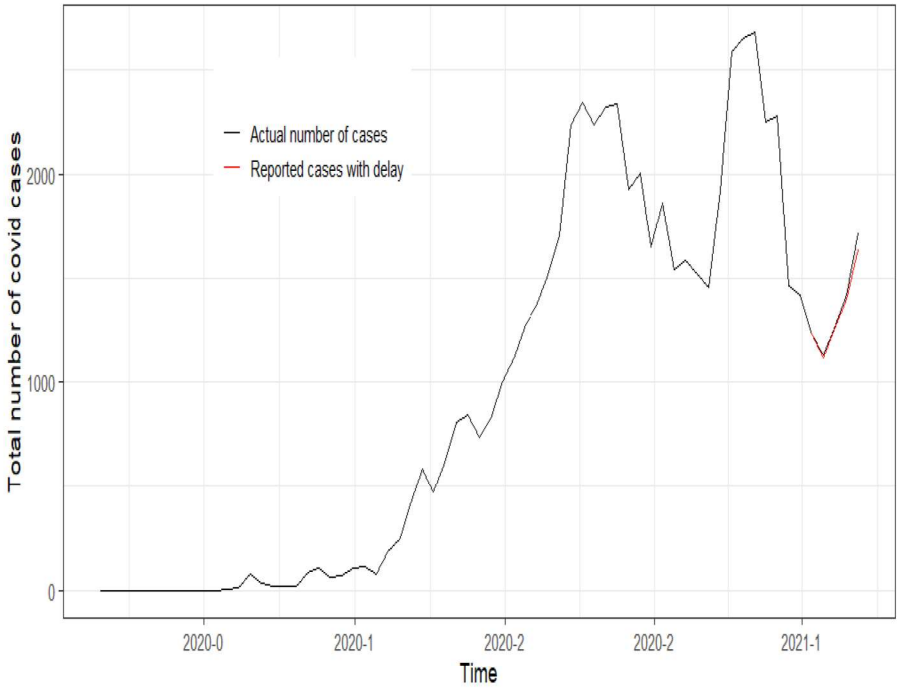

Figure 1: Total number of COVID-19 Cases from March 2020- April 2021 for the whole island of Mindanao.

Figure 1 shows the total number of COVID-19 cases from the month of March 2020 up to April 2021 for the whole island of Mindanao. The black solid line shows the actual number of cases and the red solid line shows the reported cases with delay for the last 4 weeks up to and

including the 14th pandemic week ending at April 30, 2021.

Each province is coded with a numerical value for ease of programming. Table 1 and Figure 2 presents the corresponding code of each province. Numerical coding helps avoid getting lost in large script bases, unfamiliar scripts, or legacy scripts. For example, when you're debugging, you might have to look at the provincial names across many files and projects. The use of numerical codes in the map helps to navigate around the area to easily understand the relationships between them.

Table 1: Numerical Code Identification from 1-27 for the 27 Provinces of Mindanao.

Region	Code	Province	Region	Code	Province
IX	1	Zambaonga del Norte	XII	14	Cotabato (North)
	2	Zamboanga del Sur		15	Sarangani
	3	Zamboanga Sibugay		16	South Cotabato
X	4	Bukidnon		17	Sultan Kudarat
	5	Camiguin	XIII	18	Agusan del Norte
	6	Lanao del Norte		19	Agusan del Sur
	7	Misamis Occidental		20	Dinagat Islands
	8	Misamis Oriental		21	Surigao del Norte
XI	9	Davao de Oro		22	Surigao del Sur
	10	Davao del Norte	BARMM	23	Basilan
	11	Davao del Sur		24	Lanao del Sur
	12	Davao Occidental		25	Maguindanao
	13	Davao Oriental		26	Sulu
				27	Tawi-tawi

Figure 2 is the representation of the shapefile for Mindanao, Philippines containing 27 provinces and 449 municipalities. This is loaded to the R workspace to execute the commands and algorithms of the procedure.

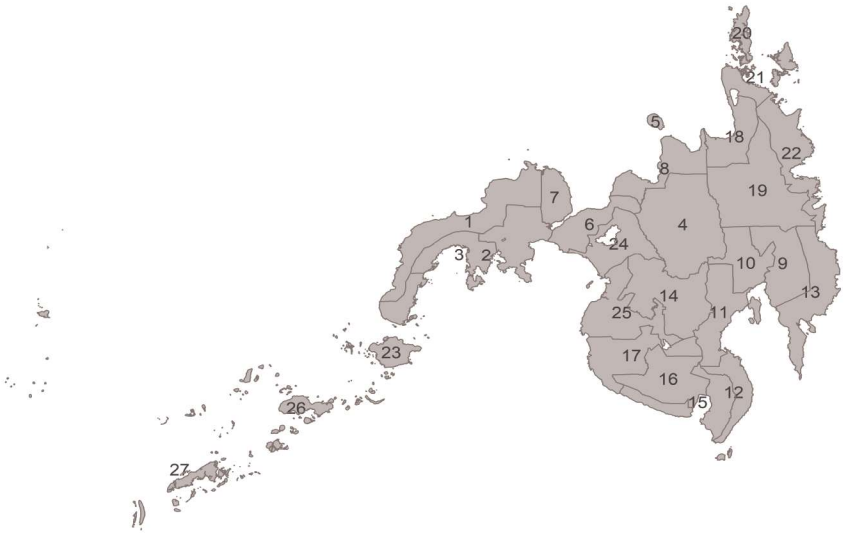

Figure 2: Provincial Map of Mindanao with Numerical Code Identification

3.2 Procedure of the Study

Figure 3 is utilized to achieve the objectives of the study.

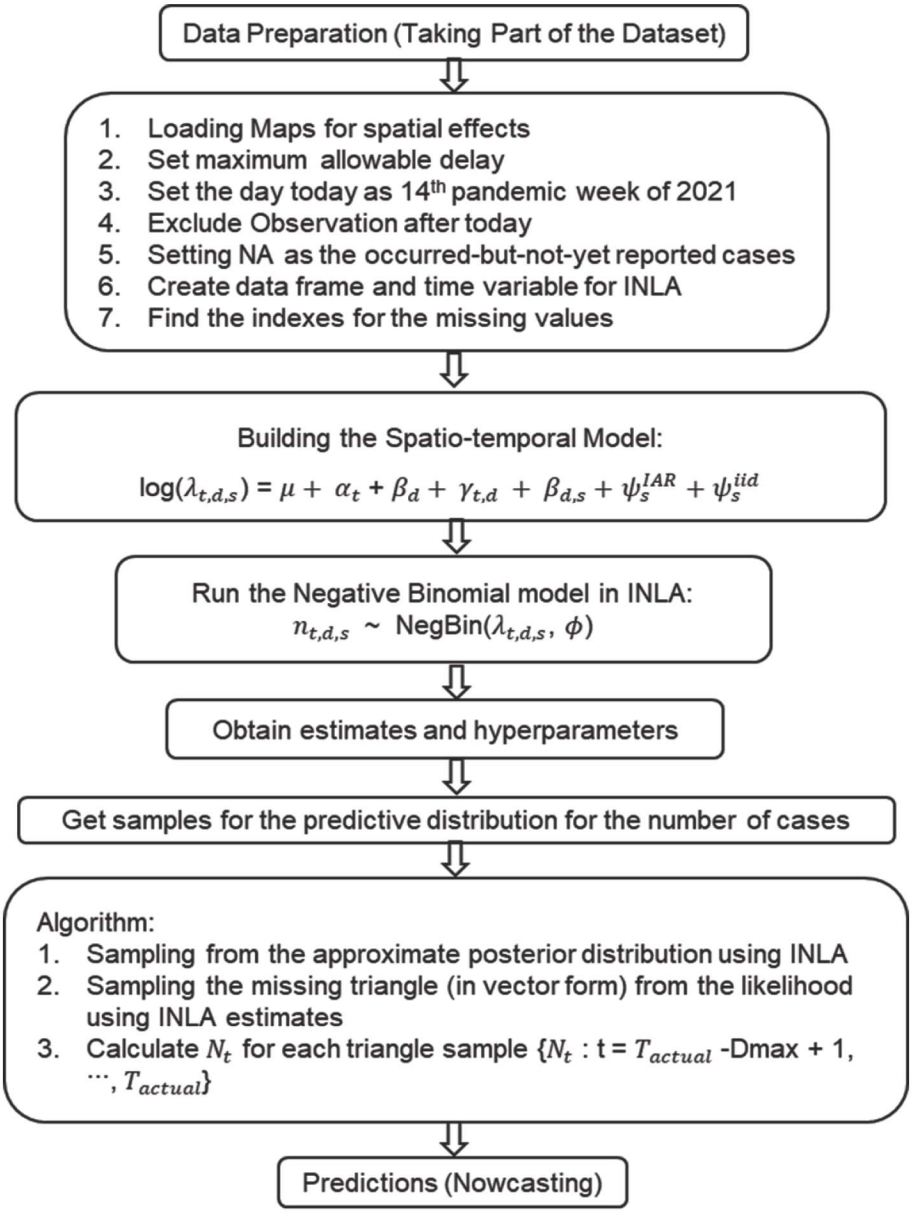

Figure 3: Schematic Diagram of the proposed Bayesian Spatio-temporal Approach to nowcasting COVID-19 cases.

3.3 Spatio-temporal Variation

In many applications, the data may be spatially grouped, for example, into a number of administrative regions spanning Mindanao. 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 analyze all data together by extending the model to allow for spatial variation not only in the process of giving rise to the counts but also in the delay mechanism. This

allows for the 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, therefore, includes spatial (Gaussian) random effects. Considering spatial variation where *s* = *S* denotes a spatial location or area in some spatial domain *s*, the model is now

$$n_{t,d,s} = NegBin(\lambda_{t,d,s}, \phi), \quad \lambda_{t,d,s} > 0, \quad \phi > 0, \tag{1}$$

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

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

where $\mathbf{X}'_{t,d,s}$ is now a model matrix that may also contain spatially varying covariates. The quantities α_t and β_d are interpreted as the overall temporal and delay evolution across space, respectively. 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 COVID-19 data, where the spatial region is fairly small so the temporal effects (α_t) are 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 [13, 4]. The spatial effect ψ_s can be defined by an intrinsic autoregressive (IAR) process [3] 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, for example, so that spatial dependence decreases exponentially with distance. In the application of the model, where space is divided into a number of administrative areas, we use the type I space-time interaction as proposed by Knorr-Held [13]. This is a formulation where

$$\beta_{d,s} \sim N(\beta_{d-1,s}, \omega^2_\beta) \tag{3}$$

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

$$\psi_s^{IAR} | \psi_{s' \neq s}^{IAR} \sim N\left(\frac{\sum_{s' \neq s} \omega_{s,s'} \psi_{s'}^{IAR}}{\sum_{s' \neq s} \omega_{s,s'}}, \frac{\sigma^2_{IAR}}{\sum_{s' \neq s} \omega_{s,s'}} \right) \tag{4}$$

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

3.4 Constructing a Parameterized Prior Distribution

Let $n_{t,d,s}$ be the notified number of cases in week t delayed in d weeks occur in region s, where $t = 1, 2, \ldots, T$, $d = 0, 1, 2, \ldots, D$, and $s = 1, 2, \ldots, 27$ provinces of Mindanao. Note that if $t + d > T$, then $n_{t,d,s}$ is unknown.

We assume a negative binomial likelihood, as the following

$$n_{t,d,s} \sim NegBin(\lambda_{t,d,s}, \phi),$$

for any $t = 1, 2, \ldots, T$, $d = 0, 1, \ldots, D$, and $s = 1, 2, \ldots, 27$. A gamma prior is set to ϕ, and the rate $\lambda_{t,d,r}$ is given by

$$\ln(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \beta'_{d,s} + \Psi_s^{(IAR)} + \Psi_s^{(ind)}.$$

A fixed effect μ was set an improper prior proportional to one. The random effects, $\{\alpha_t\}$, $\{\beta_d\}$, $\{\gamma_{t,d}\}$ were set with different random walk priors, $\{\beta'_{d,s}\}$ is an independent Gaussian space-delay random effect, and the sum $\Psi_s^{(IAR)} + \Psi_s^{(ind)}$ is model as a bym (Besag, York, and Molied) random effect, all implemented in the INLA package. And the hyperparameters, all random effects standard deviations were assumed to be half normal, or truncated normal at $(0, \infty)$, with a distinct standard deviation τ for each random effect, denoted as $HN(\tau)$. Table 2 summarizes all priors and hyperpriors for the COVID-19 model [2].

Table 2: Prior distribution for all parameters.

Parameter	Distribution	In INLA
ϕ	$\phi \sim Gamma(1, 0.1)$	$e^{\phi} \sim loggamma(1.0, 0.1)$
μ	$p(\mu) \propto 1$	default
$\alpha_t \mid \alpha_{t-1}, \sigma_\alpha^2$	$\alpha_t - \alpha_{t-1} \mid \sigma_\alpha^2 \sim N(0, \sigma_\alpha^2)$	1st order random walk (rw1)
σ_α^2	$\sigma_\alpha^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)
$\beta_d \mid \beta_{d-1}, \sigma_\beta^2$	$\beta_d - \beta_{d-1} \mid \sigma_\beta^2 \sim N(0, \sigma_\beta^2)$	1st order random walk (rw1)
σ_β^2	$\sigma_\beta^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)
$\gamma_{d,t} \mid \gamma_{d,t-1}, \sigma_\gamma^2$	$\gamma_{d,t} - \gamma_{d,t-1} \mid \sigma_\gamma^2 \sim N(0, \sigma_\gamma^2)$	1st order random walk (rw1)
σ_γ^2	$\sigma_\gamma^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)
$\beta'_{d,s} \mid \sigma_{\beta'}^2$	$\beta'_{d,s} \mid \sigma_{\beta'}^2 \sim N(0, \sigma_{\beta'}^2)$	Independent gaussian (iid)
$\sigma_{\beta'}^2$	$\sigma_{\beta'}^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)
$\Psi_s \mid \sigma_{IAR}^2, \sigma_{ind}^2$	$\Psi_s = (\Psi_s^{IAR} + \Psi_s^{ind}, \Psi_s^{ind})$	Besag-York-Mollier model
σ_{IAR}^2	$\sigma_{IAR}^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)
σ_{ind}^2	$\sigma_{ind}^2 \sim HN(\tau = 0.1)$	half_normal_sd(0.1)

By construction in INLA, the Besag-York-Mollier model [3] is a representation of an IAR model added by an unstructured independent random effect.

3.5 Parameter Estimates

In Bayesian modelling in this paper, unknown parameters are treated as random variables, each associated with a probability distribution, and are approximated by statistical models as shown in Table 3.

Table 3: Random Effects Model Descriptions

Name	**Model**
Time	RW1 model
Delay	RW1 model
Time-delay	RW1 model
Space	BYM model
Space-delay	IID model

The model described in section 3.4 is considered for parameter estimation, namely

$$n_{t,d,s} \sim NegBin(\lambda_{t,d,s}, \phi)$$

$$log(\lambda_{t,d,s}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \beta_{d,s} + \psi_s^{IAR} + \psi_s^{ind} \tag{5}$$

with t=1,$\cdots$,66(weeks), d=0,$\cdots$,D(maximum delay), and s=1,$\cdots$,27 (provinces). The quantities are again defined as,

- μ is the overall mean count at the log-scale. A fixed effect μ was set an improper prior proportional to one.
- the random effects α_t captures the mean temporal evolution of the count-generating process.
- β_d capture the mean structure of the delay mechanism. These can be modelled using random walks, in the simplest case, first-order ones.
- $\gamma_{t,d}$ is the time-delay interaction term.
- $\beta_{d,s} \sim N(\beta_{d-1,s}, \omega_\beta^2)$ allow for unstructured spatio-delay variability.
- ψ_s^{IAR} is spatially structured according to an IAR process with a neighbouring structure defined by a 27 × 27 adjacency matrix **W**, where $w_{i,j} = 1$ if the province *i* is an administrative neighbour of province *j*, and $w_{i,j} = 0$, otherwise.
- $\psi_s^{ind} \sim N(0, \sigma_\psi^2)$ captures spatially unstructured variability.

When constructing a statistical model, the Bayesian approach requires us to assign prior probability distributions (a mathematical way to reflect our prior belief) to all the unknown parameters. This gives rise to several advantages when analyzing spatio-temporal data, impacting on every aspect of statistical analysis from model building, parameter estimation, and interpretation to model evaluation.

3.6 Nowcasting

Nowcasting is defined as the process of predicting the present, the very recent past, and the very near future using time series data known to be incomplete [7]. At any given time step T, there are a number of occurred-but-not-yet-reported (missing) values $n_{\iota,d,s}$, $t = T - D + 1, \cdots, T; d = 1, \cdots, D$; $s = 1, \cdots, 27$, as well as the marginal totals $N_T - D + 1, \cdots, N_T$. The total, N_T, is obviously of primary interest and must be nowcast; however, hindcasts of $N_T - D + 1, \cdots, N_{T-1}$ may also be of interest, particularly if one wishes 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

$$p(n_{t,d,s}|\mathbf{n}) = \int_\theta p(n_{t,d}|\theta)p(\theta|\mathbf{n})d\theta, \tag{6}$$

where **n** denotes all the data used to fit the model. This cannot be solved analytically, however, with samples from the posterior $p(\theta|\mathbf{n})$ one can use Monte Carlo to approximate. In practice, for each sample from $p(\theta|\mathbf{n})$ we simulate a value from the negative binomial $p(n_{t,d}|\theta)$ to obtain an approximate sample from the predictive distribution $p(n_{t,d}|\mathbf{n})$. 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 equivalent samples can be obtained from $p(N_t)$, the marginal totals. Samples from an approximation of the joint posterior distribution $p(\theta|\mathbf{n})$ can be obtained from R-INLA using the inla.posterior.sample() function as also illustrated in small-area estimation, for example, in the work of Vandendijck et al [25].

4 Results and Discussions

4.1 Posterior Estimates

Recall that in Equation 2, given by $log(\lambda_{t,d,s})$ = μ + α_t + β_d + $\gamma_{t,d}$ + $\beta_{d,s}$ + ψ_s^{IAR} + ψ_s^{ind} + $\mathbf{X}'_{t,d,s}\delta$, the term $\mathbf{X}'_{t,d,s}$ is removed since no covariate information is available, that is, $\mathbf{X}'_{t,d,s} = 0$. Udpated equation is shown in Equation 5, that is, $log(\lambda_{t,d,s})$ = μ + α_t + β_d + $\gamma_{t,d}$ + $\beta_{d,s}$ + ψ_s^{IAR} + ψ_s^{ind}. The model assumes that the delay structure varies across provinces, through $\beta_{d,s}$ while the overall temporal evolution of the disease counts α_t is the same across the provinces. This is because the provinces are close to each other 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 $\psi_s^{IAR} + \psi_s^{ind}$ captures overall similarity in disease counts across the provinces; however, it also allows for some provinces to be different (on average) if there is such evidence in the data.

Table 4: Random Effects Precision of Posterior Estimates for the Spatio-temporal Model.

	$mean_p$	SD_p	2.5%	50%	97.5%
α_t	9.201	1.475	7.207	8.888	12.877
β_d	0.699	0.580	0.038	0.541	2.112
$\gamma_{t,d}$	9.955	1.278	8.382	9.655	13.189
$\beta_{d,s}$	4.066	1.510	2.516	3.637	8.057
ψ_s^{IAR}	36.595	27.930	10.116	28.525	110.587
ψ_s^{ind}	7.862	0.831	5.987	7.986	9.070

Table 4 presents the posterior estimates with their precision for the random effects of the Spatio-temporal model estimated by INLA. It shows the various random effects associated with its posterior mean $(mean_p)$, posterior standard error (SD_p), and uncertainty in the form 95% credible intervals with the lower and upper limits, which are 25% and 97.5%, respectively. Lastly, the posterior median estimates are in the 50% column.

4.1.1 Posterior Mean for Time Random Effects

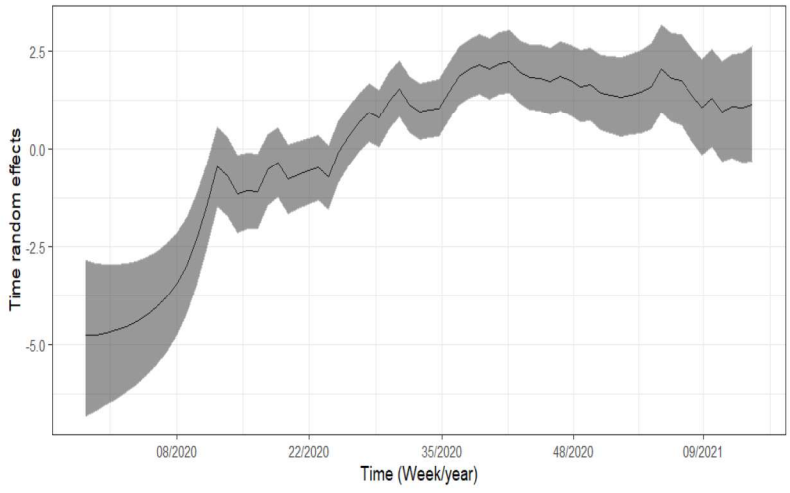

Figure 4: Posterior mean with 95% credible intervals for time on the weekly COVID-19 cases.

Figure 4 shows the posterior mean with 95% credible intervals for time random effects, α_t. The time series on the weekly COVID-19 starts from the 1st epidemic week of the year 2020 and ends at the 14th epidemic week of the year 2021.

As shown, it has a sudden increase from March to April, or about the 10th to 13th week of 2020, and a gradual increase and slight fluctuations from about the 14th week of 2020 up to the 14th week of 2021. We can also see that the number of COVID-19 cases increases as the time approaches 2021 compared to the year 2020. Moreover, the overall temporal effects do not follow any seasonal pattern.

4.1.2 Posterior Mean for Space-delay Random Effects

Posterior mean estimates of the delay mechanism, which is different across provinces, β_d + $\beta_{d,s}$, are shown in Figure 5.

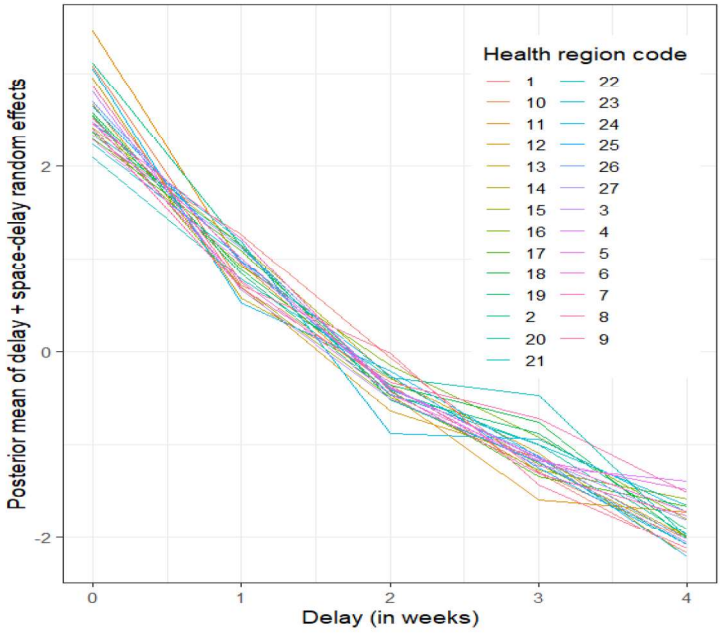

Figure 5: Posterior mean of the space-delay random effects by province.

On average, the mean reporting count decreases with delay (in weeks); however, there is considerable variability across the regions, particularly during the second and third weeks of delay. This reflects the fact that delays are likely related to several factors such as between-region differences, improvements in reporting efficiency over time, and/or weekly cycles which vary considerably in space.

4.1.3 Posterior Mean for Time-delay Random effects

Posterior mean estimates of the time-delay interaction term $\lambda_{t,d}$ where d = 0,1,2,3,4 weeks are shown in Figure 6. For d = 0 (no delay), the number of COVID-19 cases associated with the temporal evolution clearly increases. Also, the temporal evolution for d = 0 (no delay) and d = 1 (1-week delay) is negative in the first quarter of each week, suggesting that possible awareness of the epidemic leads to faster notifications or reports when a case is known.

For d = 2,3,4, respectively, the mean lies near zero in the first quarter of each week but fluctuates in later weeks up to the 14th week of 2021. It does not show an increasing trend

of delayed reports, but an inconsistent delay. This is probably due to fewer improvements in reporting efficiency over time. The delay random effect shows the importance of the delays as the number of weeks increases, and the delay should not be neglected since it has a significant effect on the real-time case notification.

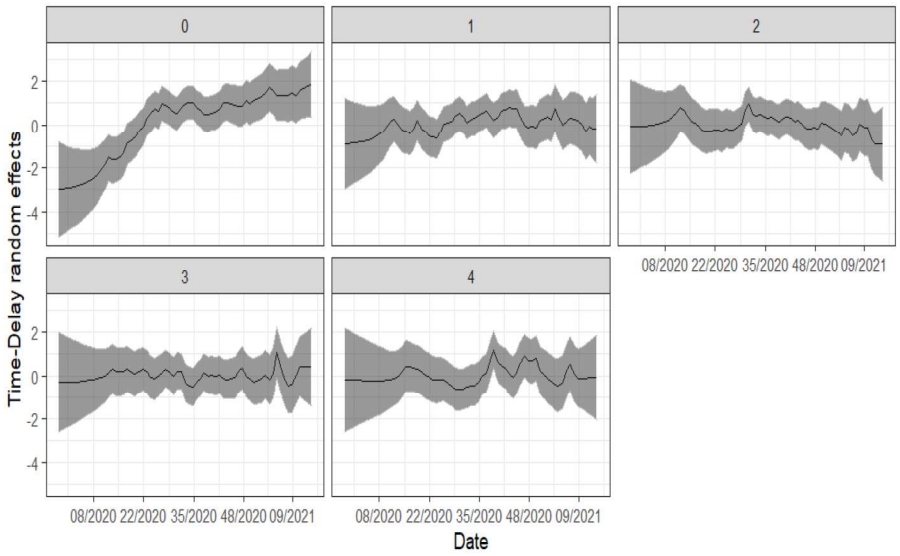

Figure 6: Posterior mean of the time-delay random effects $\lambda_{t,d}$ for $d = 0, 1, 2, 3, 4$.

4.1.4 Posterior Mean for the Overall Spatial Random Effects

Posterior mean estimate of the overall spatial variability term, $\Psi_s^{(IAR)} + \Psi_s^{(ind)}$, is shown in Figure 7. This indicates some variability in the number of COVID-19 reports across the provinces, but also a similarity in neighboring regions. This is probably reflecting unobserved factors relating to the susceptible population (including population size). The delay can vary from place to place, being susceptible to the adherence of health care providers to the notification protocol as well as the access of patients to health care and health system shortcomings.

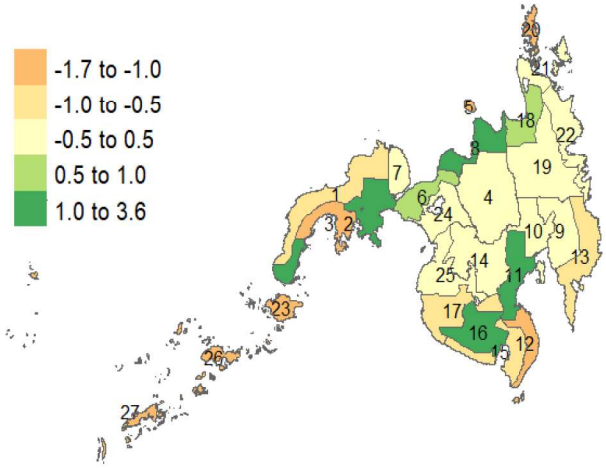

Figure 7: Posterior mean of the spatial random effects ψ_s.

4.1.5 Posterior Distribution of the Spatial Correlation

In order to assess whether the spatial correlation was adequately captured, we consider the measure

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

This quantifies the contribution of the structured random effect ψ_s^{IAR} to the total variance of the spatial effect $\psi_s^{IAR} + \psi_s^{ind}$.

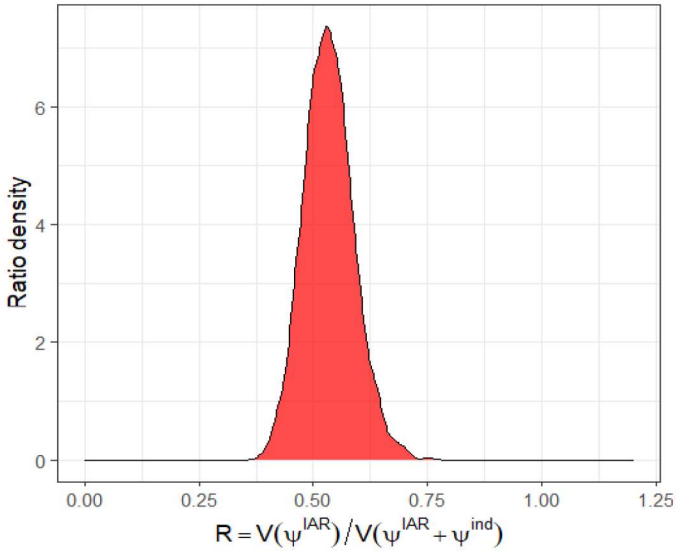

Figure 8: Posterior distribution of R

Values close to zero indicate that there is little spatial correlation, whereas values around 0.5 indicate that structured and unstructured spatial effects contribute roughly equally. 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 variability.

4.2 Nowcasting Results

The goal of this study is to use the proposed model to correct reporting delays by considering spatial and temporal variability in the delay mechanism. The proposed delay model was implemented without covariates, and the predicted cases were estimated using their predictive posterior distribution. As previously explained, our model can be used to directly $nowcast$ or correct reports from previous days for the delay.

Given that we model delay d as a stand-alone variable in our Spatio-temporal model, we are able to predict the missing cells directly by setting the delay d to the necessary value in the data vector used for predictions.

Figure 9 shows the time series of reported COVID-19 cases, as well as predictions or $nowcast\ values$ in the whole island of Mindanao as the maximum possible delay, which is set to 4 weeks. Subfigures A and B differ only on the time scale, where A starts from January 2020, whereas B starts from January 2021. It also depicts the estimated mean as well as the 95% prediction intervals (dotted black line and shaded region) of the corresponding predictive distribution from Equation 5, that is, $n_{t,d,s} \sim NegBin(\lambda_{t,d,s}, \phi)$, where $log(\lambda_{t,d,s})$ = μ + α_t + β_d + $\gamma_{t,d}$ + $\beta_{d,s}$ + ψ_s^{IAR} + ψ_s^{ind} .

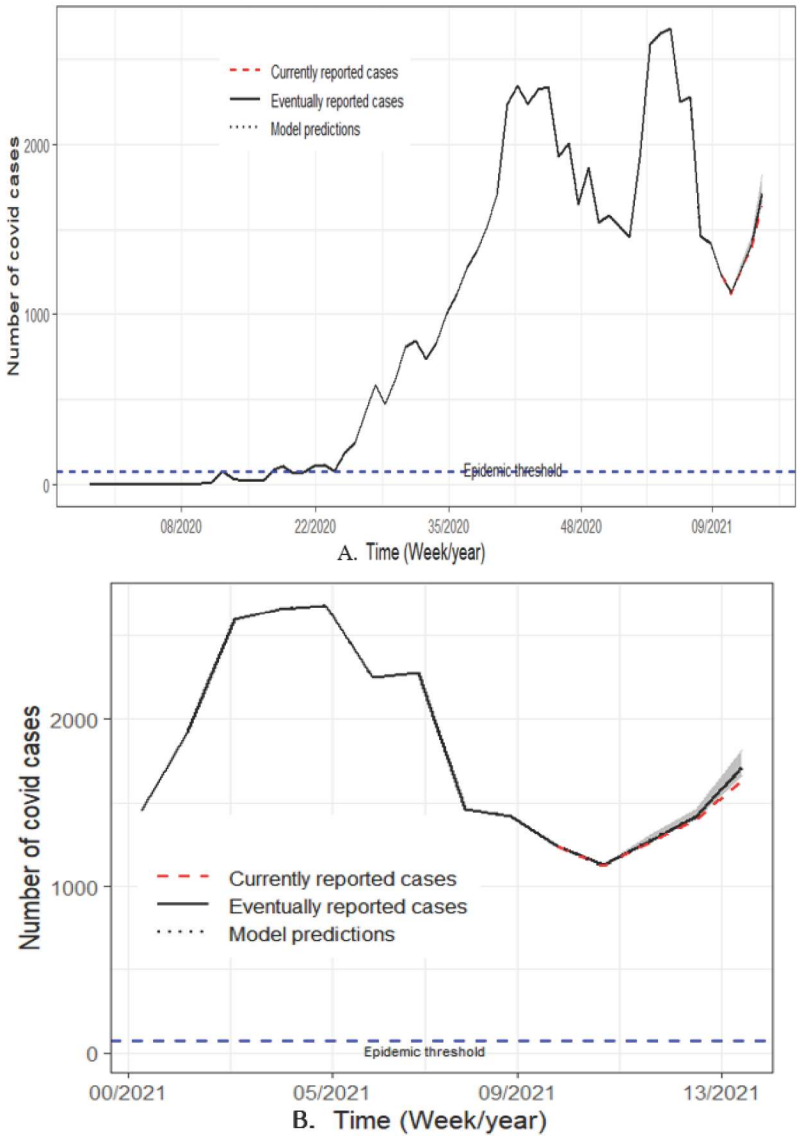

Figure 9: Time series of COVID-19 cases in the whole island of Mindanao, Philippines.

The plot also shows the weekly time series of the eventually reported COVID-19 cases in Mindanao, Philippines from the first epidemic week of 2020 to the 14th epidemic week of 2021 (solid black line) in subfigure A and COVID-19 cases from the first epidemic week of 2021 to the 14th epidemic week of 2021 (solid black line) in subfigure B. Finally, the plot also shows the currently reported number of COVID-19 cases for the last 4 weeks, up to and including the 14th epidemic week ending on April 2, 2017 (dashed red line).

In week 14, the currently reported number of cases is $n_{14,0} = 1633$, and the eventually reported count is $\sum_{d=0}^{4} n_{14,d} = 1710$, while the predicted value for N_t is 1708. Consequently, 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).

Furthermore, the amount of error in the model is measured using the mean square error (MSE) to assess the average squared difference between the actual or eventually reported cases and the predicted values or model predictions, as shown in Figure 9 . In other words,

$MSE = 1/n \sum_1^n (Y_i - \hat{Y}_i)^2$. The computed MSE is 4.0, which is small, indicating a more precise prediction.

Another test to measure the amount of error is the mean absolute deviation (MAD). It is used to assess the average distance between each data point in the predicted values or $nowcast\ values$ and its mean. The computed MAD is 219.424 (high) and tells us that many of the data values are spread out further from the mean.

Finally, nowcasting estimation is of utmost importance in order to provide accurate and reliable estimations to avoid misclassification of warning issuance [2]. These disease-specific quantities provide a tool for setting goals for reporting delays, not only for outbreak control but also for evaluation of individual-based interventions with other aims, such as partially reducing infections or completely stopping transmission.

5 Conclusion and Recommendation

In disease surveillance, the spatial and temporal components in a pandemic or disease outbreaks are essential so that the strength, direction, and trend of the disease transmission are considered. The necessity in a disease outbreak lies in predicting or $nowcasting$ the total number of disease cases, $n_{t,d,s}$, to aid health authorities to have effective control measures and issuance of warnings to the public. The Bayesian hierarchical framework is implemented in R-INLA to explore the possibility of a flexible way of correcting delayed reports considering time and space. Results show that the proposed model was able to capture the increasing trend of COVID-19 disease counts in the presence of delayed reports. The data, however, do not contain any information on covariates. Therefore, the author advises using the modelling technique on epidemiological data that contains confounders.

6 Acknowledgements

Jejemae D. Nacion is grateful to the DOST-SEI ASTHRDP for the financial support during her MS Statistics degree program at MSU-IIT.

References

[1] Bacchetti, P., M. R. Segal, and N. P. Jewell. 1993. "Backcalculation of HIV infection rates." Statistical Science 8 (2), 82–101.

[2] Bastos, Leonardo S et al. 2019. "A Modelling Approach for Correcting Reporting Delays in Disease Surveillance Data." Statistics in Medicine 38(22): 4363–77.

[3] Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." Annals of the Institute of Statistical Mathematics 43(1): 1–20.

[4] Blangiardo, Marta, Michela Cameletti, Gianluca Baio, and Håvard Rue. 2013. "Spatial and Spatio-Temporal Models with R-INLA." Spatial and Spatio-Temporal Epidemiology 4: 33–49.

[5] Brookmeyer, R., and A. Damiano. 1989. "Statistical Methods for Short-Term Projections of AIDS Incidence." Statistics in Medicine 8(1): 23–34.

[6] Brookmeyer, Ron, and Mitchell H. Gail. 1988. "A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the Aids Epidemic." Journal of the American Statistical Association 83(402): 301–8.

[7] Giannone, Domenico, Lucrezia Reichlin, and Marta Bańbura. 2010. Working Paper Series Nowcasting. European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20101275.html (July 13, 2022).

[8] Harris, J. E. 1990. "Reporting delays and the incidence of AIDS." Journal of the American Statistical Association 85 (412), 915–924.

[9] Höhle, Michael, and Matthias an der Heiden. 2014. "Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011." Biometrics 70.

[10] Kalbfleisch, J. D., and J. F. Lawless. 1989. "Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS." Journal of the American Statistical Association 84(406): 360–72.

[11] Kalbfleisch, J. D. and J. F. Lawless 1991. "Regression models for right truncated data with applications to AIDS incubation times and reporting lags." Statistica Sinica 1 (1), 19–32.

[12] van de Kassteele, Jan, Paul H. C. Eilers, and Jacco Wallinga. 2019. "Nowcasting the Number of New Symptomatic Cases During Infectious Disease Outbreaks Using Constrained P-Spline Smoothing." Epidemiology (Cambridge, Mass.) 30(5): 737–45.

[13] Knorr-Held, L. 2000. "Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk." Statistics in Medicine 19(17–18): 2555–67.

[14] Klaucke DN, Buehler JW, et.al. [1988] Guidelines for evaluating surveillance systems. Morb Mortal Wkly Rep. 37(Suppl5):1-18.

[15] Kline, David et al. 2021. "A Bayesian Spatio-Temporal Nowcasting Model for Public Health Decision-Making and Surveillance." https://arxiv.org/abs/2102.04544v1 (February 16, 2022).

[16] Lagakos, S., L. Barraj, and V. Gruttola. 1988. "Nonparametric Analysis of Truncated Survival Data, with Application to AIDS."

[17] Lawless, J. F. 1994. “Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events.” The Canadian Journal of Statistics / La Revue Canadienne de Statistique 22(1): 15–31.

[18] McGough, Sarah F., Michael A. Johansson, Marc Lipsitch, and Nicolas A. Menzies. 2020. “Nowcasting by Bayesian Smoothing: A Flexible, Generalizable Model for Real-Time Epidemic Tracking.” PLOS Computational Biology 16(4): e1007735.

[19] Renshaw, A. E., and R. J. Verrall. 1998. “A Stochastic Model Underlying the Chain-Ladder Technique.” British Actuarial Journal 4(4): 903–23.

[20] Rotejanaprasert, Chawarat, Nattwut Ekapirat, Darin Areechokchai, and Richard J. Maude. 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): 4.

[21] Salmon, M., D. Schumacher, K. Stark, and M. Höhle 2015. ”Bayesian outbreak detection in the presence of reporting delays.” Biometrical Journal 57(6), 1051–1067.

[22] Stoner, Oliver, and Theo Economou. 2020. “Multivariate Hierarchical Frameworks for Modeling Delayed Reporting in Count Data.” Biometrics 76(3): 789–98.

[23] Stoner, Oliver, and Theodoros Economou. 2019. ”A Hierarchical Modelling Framework for Correcting Delayed Reporting in Spatio-Temporal Disease Surveillance Data.” arXiv. 2019.

[24] Swaan, Corien, Anouk van den Broek, Mirjam Kretzschmar, and Jan Hendrik Richardus. 2018. “Timeliness of Notification Systems for Infectious Diseases: A Systematic Literature Review.” PLOS ONE 13(6): e0198845.

[25] Vandendijck, Y. et al. 2016. “Model-Based Inference for Small Area Estimation with Sampling Weights.” Spatial Statistics 18: 455–73.