Hierarchical Bayesian Model for Correcting Reporting Delays in Dengue Counts[1]

Mikee T. Demecillo[2] and Bernadette F. Tubo
Department of Mathematics and Statistics
MSU-Iligan Institute of Technology Iligan City, Philippines

ABSTRACT

Real-time surveillance and precise case estimation are necessary for situational awareness in order to spot trends and outbreaks and establish efficient control actions. The comprehension of the mechanisms of a sudden rise or fall in disease cases that change over time is hampered by the reporting delays between disease start and case reporting. This study uses a flexible temporal nowcasting model with a Bayesian inference for latent Gaussian models built in R-INLA to rectify reporting delays for weekly dengue surveillance data in Northern Mindanao from 2009 to 2010. Additionally, it seeks to quantify all the uncertainties involved in replacing the missing value. The statistical issue is to forecast run-off triangle numbers based on actual counts $n_{t,d}$. In contrast to the currently reported instances, which seem to be declining, the posterior predictive model on the given temporal dataset recognizes the fact that there are more dengue cases than there were previously (supporting the actual scenario). This implies that even with delayed data, the model was still able to provide a reliable estimate of the true number of instances. This paper offers a model for nowcasting to aid in dengue control and good judgment on the part of interested authorities.

Keywords *– Latent Gaussian Model, Nowcast, Count Data*

I. INTRODUCTION

Epidemiological surveillance is the systematic collection, analysis, and dissemination of health data for public health purposes according to Klaucke, et al. (1988). Identifying outbreaks and launching prompt response is one of the duties of infectious disease surveillance. In many applications from this discipline, count data are generated that might not accurately reflect the quantity of interest. In accord with Farrington, et al. (1996), detecting an early increase in diseases is frequently difficult since reports must be examined and acted on as they accumulate, with little opportunity to correct errors or compensate for reporting delays and other reporting system abnormalities.

Infectious disease control requires effective and prompt responses to unanticipated increases in disease burden. As reported, public health control strategies may be greatly impacted by an inability to generate a timely and accurate estimation of the burden of contagious diseases. A virus borne by mosquitoes called dengue has this type of issue with

[1] Presented at the 15th National Convention on Statistics, 03-5 October, 2022, Crowne Plaza Manila Galleria, Ortigas Center, Quezon City and published in this issue as a *Technical Paper*.
[2] Corresponding author: *mikee.demecillo@g.msuiit.edu.ph*

tropical nations as one of the diseases that is affected by delay problems. A significant challenge for dengue monitoring is the time lag between the start of adverse health events and reporting, as well as the time lag between reporting and the identification of trends or outbreaks. This causes a delayed or nonexistent response as well as an initial underestimation of the true scenario. It lengthens the time required to respond to catastrophic outbreaks and puts lives at risk.

A timely and accurate disease count, with no delays, is vital for monitoring health outcome trends and detecting disease outbreaks that vary over time. Delay reporting, on the other hand, is where the total observable count, which may still be fewer than the true count, is only available after a specific length of time. As a result, it is a major issue when decisions are made based on total counts required before everything has been thoroughly inspected. As a result, forecasts regarding the current condition of the disease must be made based on partial counts observed, in order to monitor and detect abrupt spikes that may necessitate prompt public health responses and vector control measures.

Rosinska, et al., (2015) stated that the problem of occurred-but-not-yet-reported cases during outbreaks is well known from the HIV/AIDS outbreak, and different statistical approaches have been proposed to handle delayed reporting. A standard reference is Lawless (1994). However, a more flexible Bayesian nowcasting approach has been developed by Höhle, et al. (2014). By allowing for temporal fluctuation in both the overall number of cases N_t and the delay mechanism, the framework was used to address delays. The reporting delay is defined as the interval between the onset and the official case reporting by a health authority, where the delay correction approach is referred to as the nowcast. Additionally, Bastos, et al. (2019) suggested utilizing a lognormal survival model in conjunction with a Bayesian hierarchical modeling technique to rectify reporting delay and associated uncertainty. They apply this paradigm to data on spatiotemporal severe acute respiratory infections (SARI) in the state of Paraná (Brazil) and dengue fever in Rio de Janeiro using integrated nested Laplace approximations (INLA). The marginal distribution of each count $n_{t,d,s}$ is negative binomial with mean $\lambda_{t,d,s}$ and dispersion parameter, $\emptyset$ to allow spatiotemporal variation in counts, as well as covariates.

Inspired by the work of Bastos, et al. (2019), provides a decision-support tool that examines the Bayesian hierarchical model for count data with time delays (temporal model), which is sufficiently adaptable to be used for a wide range of clinical applications. The suggested model, a latent Gaussian, is readily implemented using R-INLA, which roughly approximated the marginal posterior distributions of the latent fields and the hyperparameters. This study also considers the effects of delays that occur in reporting events such as cases of a reportable disease like dengue cases in Northern Mindanao from 2009 to 2010 and in estimating the number of events that have *occurred but not yet reported* (OBNR) events for temporal model.

II. RELATED LITERATURE

Historically, as reported by Renshaw, et al., (1998), actuarial sciences have accounted for systematic reporting delays when modeling claims reserves. It has been addressed for HIV/AIDS-related health outcomes of Brookmeyer (1989), mortality reporting Lin, et al., (2008), and chronic diseases, including cancer registries, Midthune (2005). Generally, in accordance with Brookmeyer (1989), the process of rectifying delayed reporting has been isolated from the task of predicting or forecasting the overall incidence. However, this disregards the joint uncertainty in the total count incidence and the presence of delay. Suppose, for instance, that at time t the number of cases reported in the first week is typically low, $n_{t,1}$. This could be due to a low proportion of N_t reported in the first week, an exceptionally low level of N_t, or to both. Differentiating between both scenarios is essential for accurate forecasting, hence we focus on a technique that predicts both the delay mechanism and the total count.

Additionally, methods have been created to combat infectious illness outbreaks. Methods have been developed to nowcast (i.e. estimate in real-time) the current number of affected individuals. Hohle and Heiden (2014) predicted the daily number of hemolytic uremic syndrome hospitalizations. They evaluate the distribution of the counts $n_{t,d}$ with respect to the totals N_t. The framework is then hierarchical, with N_t assumed to have a Poisson or Negative Binomial distribution. $n_{t,k}|N_t$ is then multinomial with a probability vector of size D that must be calculated. Hohle, et al., (2011) suggested that the concept was incorporated into a Bayesian nowcasting model to account for reporting delays of Shiga toxin-producing *Escherichia coli* in Germany. By describing the multinomial probability vector as a function of time, the model permits smooth changes in the temporal variation of the total number of cases N_t and the delay mechanism. Then, Noufaily et al. (2013, 2016) proposed a quasi-Poisson algorithm-based technique for predicting infectious disease outbreaks from laboratory data with reporting delays.

In 1993, Mack developed the so-called chain-ladder technique as a distribution-free method to estimate missing delayed counts. The chain ladder method is likely the most commonly used technique for assessing IBNR claims reserves. Renshaw and Verral (1998) demonstrated that the model underlying the chain-ladder technique is a generalized linear model for $n_{t,d}$ where the mean is denoted by $E\left[n_{t,d}\right] = \mu + \alpha + \beta_d$. This enables the approach to process negative incremental claims, as demonstrated.

Salmon, et al., (2015) demonstrated that the conditional multinomial method might be used to motivate the chain ladder structure. Assume initially that the total counts N_t are governed by a negative binomial distribution with a certain mean λ_t and dispersion parameter ϕ; $N_t|\lambda_t, \phi \sim NB(\lambda_t, \phi)$. This is a common assumption when modeling disease count data, where the negative binomial extends the Poisson to account for overdispersion in data where the amount of susceptible population is unknown, which is a typical issue in observational monitoring data. The marginal distribution of each $n_{t,d}$ is a negative binomial with mean $\pi_{t,d}\lambda_t$ and dispersion parameter ϕ, provided the counts in each row are conditionally multinomial $n_t \sim \mathrm{MN}(\pi_t, N_t)$. In this way, the conditional multinomial technique can be used to justify the chain ladder method, which explicitly models the marginals as negative binomial. However, $\phi_{t,d}$ and λ_t cannot be separated.

The correction of time delays, as proposed by Bastos et al., (2019) will be the focus of this study. Bayesian inference for the latent Gaussian model, which is easily implemented in the INLA package, will be used. The posterior marginal of the latent Gaussian fields and the posterior marginal of the hyperparameters under investigation are both considered in this work. By using the model in the R software, its usefulness may be evaluated.

The field of infectious disease modelling includes a vast literature of methodologies, but the difficulty of estimating spatiotemporal reporting delays in real-time public health control applications has not been previously addressed. Bastos, et al., (2019) established an integrated dengue monitoring system in Rio de Janeiro, Brazil, which uses a lognormal survival model to adjust reporting delays. It assumes that the distribution of the counts $n_{t,d,s}$ follows a conditionally independent negative binomial function with the mean $\lambda_{t,d,s}$ and the scale parameter ϕ. They consider approximate Bayesian Inference in a popular subset of structured additive regression models, latent Gaussian models, where the latent fields are Gaussian, controlled by a few hyperparameters and with non-Gaussian response variables. The integrated nested Laplace approximation is used, which makes the model's implementation fast. They contrasted the spatial version model (spatiotemporal), which assumes spatial variability and dependence on borrowing information across the spatial units, with the nonspatial version model (temporal), which has dependency structures in both time and delay. By Codeco, et al., (2019), Brazilian authorities are using the model as a decision-making tool after further development and as warning systems, *infoDengue* and *infoGripe*.

III. RESEARCH METHODOLOGY

A. Data Description

An infectious disease spread by mosquitoes called dengue places a heavy impact on the economy and public health in tropical areas. For instance, according to Undurraga, et al., (2017), dengue poses a significant burden in the Philippines, over ten times more than the estimated burdens of rabies, intestinal fluke diseases, and tuberculosis combined.

The Gregorio T. Lluch Memorial Hospital in Pala-o, Iligan City, served as the source for collecting data on recorded dengue cases from the Gregorio T. Lluch Memorial Hospital in Pala-o, Iligan City. The statistics acquired to cover the time period beginning in January 2009 and ending in December 2010 and pertain to dengue cases in Northern Mindanao. A substantial number of dengue cases were reported during the first six months of 2010. The purpose of this study is to represent behavior that can exhibit periods of very low activity as well as rapid increases and decreases. In order to demonstrate this, 77 weeks spanning from 2009 to 2010 are used; 52 weeks in 2009 and the 1st to 25th weeks (June 2010). This shows an immediate spike in the number of dengue cases in Northern Mindanao.

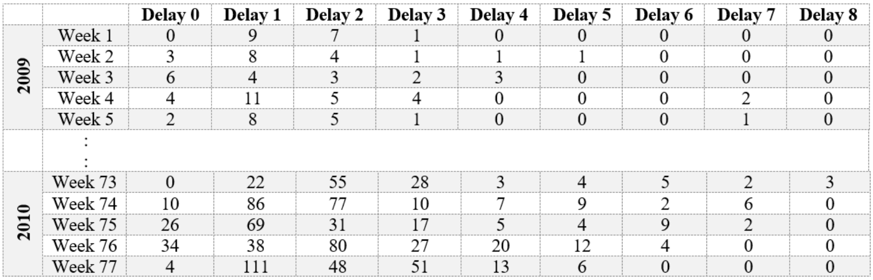
B. Data Management and Processing

Let the delay be denoted by $d = 0,1,\dots,D$ as used in the model. To calculate the delay, records must include two dates: the date of notification (date of admission) and the date of digitization (when the information is fed into the system or date of entry). Below are the procedures on how to calculate delays:

- First, subtract the date of notification from the date of digitization.
- To have weekly data, divide the obtained days of delay by seven.

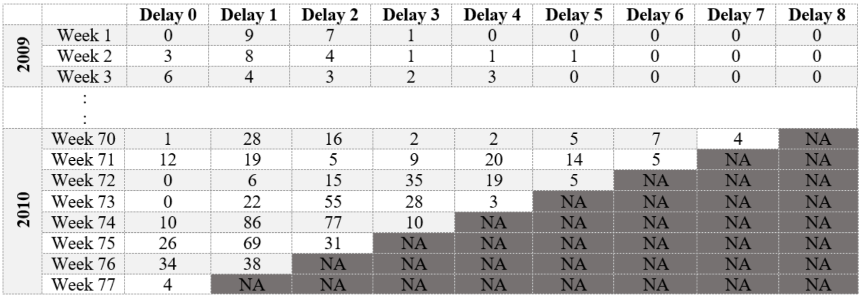
After that, remove the observations after 77 weeks. Dengue fever should be reported within seven days of the diagnosis. In actuality, however, less than 50% of applicants are notified within one week, less than 75% within four weeks, and no more than 90% within seven weeks. As a result, eight weeks is a plausible upper bound for the delay period D. Table 1 depicts the data structure of the surveillance data with reporting delays of dengue cases in Northern Mindanao.

Table 1: Time-delay data of dengue cases in Northern Mindanao.



		Delay 0	Delay 1	Delay 2	Delay 3	Delay 4	Delay 5	Delay 6	Delay 7	Delay 8
2009	Week 1	0	9	7	1	0	0	0	0	0
	Week 2	3	8	4	1	1	1	0	0	0
	Week 3	6	4	3	2	3	0	0	0	0
	Week 4	4	11	5	4	0	0	0	2	0
	Week 5	2	8	5	1	0	0	0	1	0
	: :									
2010	Week 73	0	22	55	28	3	4	5	2	3
	Week 74	10	86	77	10	7	9	2	6	0
	Week 75	26	69	31	17	5	4	9	2	0
	Week 76	34	38	80	27	20	12	4	0	0
	Week 77	4	111	48	51	13	6	0	0	0

Table 2: Data Structure for Dengue Cases with Missing Values.



		Delay 0	Delay 1	Delay 2	Delay 3	Delay 4	Delay 5	Delay 6	Delay 7	Delay 8
2009	Week 1	0	9	7	1	0	0	0	0	0
	Week 2	3	8	4	1	1	1	0	0	0
	Week 3	6	4	3	2	3	0	0	0	0
	: :									
2010	Week 70	1	28	16	2	2	5	7	4	NA
	Week 71	12	19	5	9	20	14	5	NA	NA
	Week 72	0	6	15	35	19	5	NA	NA	NA
	Week 73	0	22	55	28	3	NA	NA	NA	NA
	Week 74	10	86	77	10	NA	NA	NA	NA	NA
	Week 75	26	69	31	NA	NA	NA	NA	NA	NA
	Week 76	34	38	NA	NA	NA	NA	NA	NA	NA
	Week 77	4	NA	NA	NA	NA	NA	NA	NA	NA

Then, train the data or construct the run-off triangle data frame shown in Table 2. The table displays the data format for dengue cases, with missing values (grey) representing delayed counts. Figure 1 depicts the methodology employed to fulfill the study's purpose.

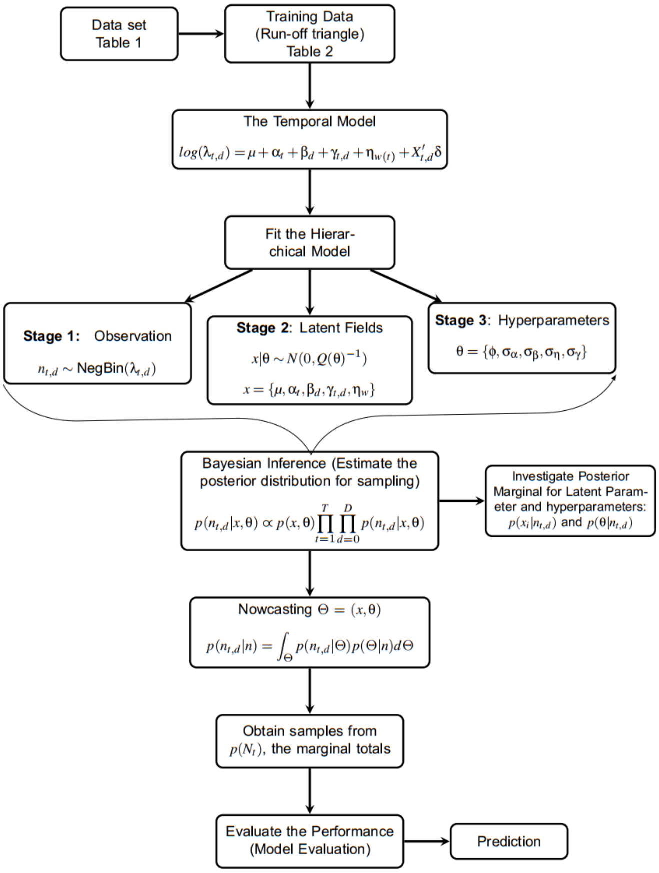

Data set Table 1
Training Data (Run-off triangle) Table 2
The Temporal Model
$log(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)} + X'_{t,d}\delta$
Fit the Hierarchical Model
Stage 1: Observation
$n_{t,d} \sim$ NegBin$(\lambda_{t,d})$
Stage 2: Latent Fields
$x|\theta \sim N(0, Q(\theta)^{-1})$
$x = \{\mu, \alpha_t, \beta_d, \gamma_{t,d}, \eta_w\}$
Stage 3: Hyperparameters
$\theta = \{\phi, \sigma_\alpha, \sigma_\beta, \sigma_\eta, \sigma_\gamma\}$
Bayesian Inference (Estimate the posterior distribution for sampling)
$p(n_{t,d}|x,\theta) \propto p(x,\theta) \prod_{t=1}^{T} \prod_{d=0}^{D} p(n_{t,d}|x,\theta)$
Investigate Posterior Marginal for Latent Parameter and hyperparameters: $p(x_i|n_{t,d})$ and $p(\theta|n_{t,d})$
Nowcasting $\Theta = (x,\theta)$
$p(n_{t,d}|n) = \int_\Theta p(n_{t,d}|\Theta) p(\Theta|n) d\Theta$
Obtain samples from $p(N_t)$, the marginal totals
Evaluate the Performance (Model Evaluation)
Prediction

Fig. 1 Schematic diagram of the proposed hierarchical Bayesian temporal model to correct delays.

C. Latent Gaussian Method

Let $n_{t,d}$ represent the number of instances reported in week t that were delayed by d weeks, where $t = 1,2,\ldots,T$ and $d = 0,1,\ldots,D$. T is the most recent time step for which data is available, and D is the greatest permissible delay, which for disease applications is potentially infinite but is assumed to be finite for the sake of simplicity. Note that if $t + d > T$, $n_{t,d}$ is occur-but-not-yet-reported and so unknown.

The INLA framework was created to handle latent Gaussian models, in which the counts $n_{t,d}$ follow a conditionally independent negative binomial probability distribution, with mean λ_t and scaling parameter ϕ , i.e.

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

The parameterization used here is such that $E[n_{t,d}] = \lambda_{t,d}$ and $V[n_{t,d}] = \lambda_{t,d}(1 + \lambda_{t,d}/\phi)$. Using a Bayesian technique, the predictive distribution of $n_{t,d}$ for any t and d (given the data) is readily available, as well as the uncertainty associated in their estimation.

The parameter $\lambda_{t,d}$ is linked to a structured additive predictor via the logarithm of their mean, $\lambda_{t,d}$, $log(\lambda_{t,d})$. This is done in order to capture the structured temporal variation in $n_{t,d}$, which is defined as follows:

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

- μ is the log-scale total mean count. An improper prior proportional to one was applied to a fixed effect μ.
- The mean temporal evolution of the count-generating process is captured by the random effects α_t.
- The mean structure of the delay mechanism is captured by β_d. These can be modeled using random walks, particularly first-order ones, i.e,

$$\alpha_t \sim N(\alpha_{t-1}, \sigma_\alpha^2), \qquad t = 2,3, \dots, T \text{ and} \tag{3}$$

$$\beta_d \sim N(\beta_{d-1}, \sigma_\beta^2), \qquad d = 1,2, \dots, D \tag{4}$$

  where half normal HN(τ^2) prior distributions are assumed for σ_α and β_d. These are distribution on [0, ∞) 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 sence "smooth") the functions will be (i.e, the smaller the first-order differences will be).
- The time-delay interaction term $\gamma_{t,d}$ is modelled as

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

  so that there is an independent realization 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.
- $\eta_{w(t)}$, where $w(t) = 1, \dots, 52$ is the week index, is a seasonal component defined as a second-order random effect,

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

  constrained in such a way that week 1 and week 52 are joined.
- $X'_{t,d}$ is a matrix of temporal and delay-related covariates with associated vector of parameters δ.

Note that 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.

The hierarchical model is then completed with an approximate prior distribution for the hyperparameters of the model $\theta = \left(\phi, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\eta^2\right)$, where $\phi \sim Gamma(1{,}0.1), \sigma_\alpha^2 \sim HN(\tau = 0.1), \sigma_\beta^2 \sim HN(\tau = 1), \sigma_\eta^2 \sim HN(\tau = 1), \sigma_\gamma^2 \sim HN(\tau = 0.1)$. Assuming an 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. A gamma prior is then set to ϕ.

The Bayesian inference for the latent Gaussian model is used to obtain a posterior marginal distribution. The joint posterior distribution for $\Theta = \left(\mu, \alpha_t, \beta_d, \gamma_{t,d}, \eta_w, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\eta^2, \sigma_\gamma^2, \phi\right)$, given all the observed data $\boldsymbol{n} = n_{t,d}$ is given by:

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

where $p\left(n_{t,d}\middle|\Theta\right)$ is negative binomial density function (1), and $p(\Theta)$ is the joint prior distribution given by the product of the prior distributions for $\phi, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\eta^2$ and the random effects distribution. To investigate the posterior marginal distribution of hyperparameters and latent Gaussian models, set the latent Gaussian vector x, $\boldsymbol{x} = \{\mu, \boldsymbol{\alpha_t}, \boldsymbol{\beta_d}, \boldsymbol{\gamma_{t,d}}, \boldsymbol{\eta_w}\}$ and hyperparameter vector θ, $\theta = \left(\phi, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\eta^2\right)$. Now, compute from

$$p(x, \theta|n) \propto p(x, \theta) \prod_{t=1}^{T} \prod_{d=0}^{D} p\left(n_{t,d}\middle|x, \theta\right), \tag{8}$$

the posterior marginal $p\left(x_i\middle|n_{t,d}\right)$, for some i; and $p\left(\theta_i\middle|n_{t,d}\right)$, for some i.

In any given step T, there are a number of occurred-but-not-yet-reported (missing) values $n_{t,d}, t = T - D + 1, \dots, T; d = 1, \dots, D$ (see the grey cells in Table 2), as well as the marginal totals $N_{T-D+1}, \dots, N_T$. Of primary interest is of course N_T, which needs to be nowcasts; however, hindcasts of $N_{T-D+1}, \dots, N_T - 1$ may also be of interest, especially if one wants to quantify the rate of increase or decrease in the counts.

From a Bayesian standpoint, this is a prediction problem in which the posterior predictive distribution can be used to estimate all of the missing $n_{t,d}$

$$p\left(n_{t,d}\middle|n\right) = \int_\Theta p\left(n_{t,d}\middle|\Theta\right) p(\Theta|n) d \tag{9}$$

where $\boldsymbol{n}$ denotes all the data used to fit the model.

IV. RESULTS AND DISCUSSION

A. Time Series of Reported Dengue Cases

The available data consist of weekly counts of the number of dengue cases in Northern Mindanao for the time period January 2009 to June 2010, along with the associated delay information. Figure 2 shows the eventually reported number of dengue cases per week as a solid dark blue line. The dashed red line represents the current reported number of cases from the 18th to the 25th week of 2010. (circled cross). The green dotted line depicts the model estimates for this period, along with the blue 95 prediction intervals.

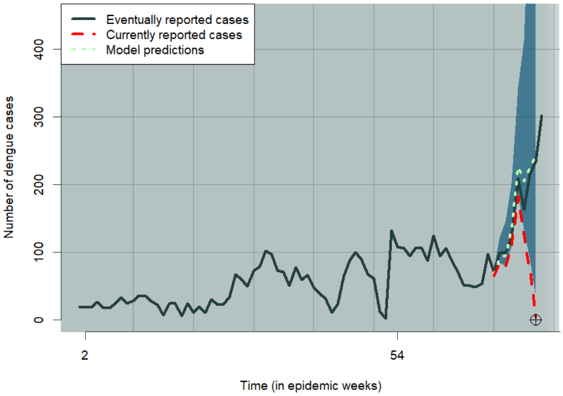


Fig. 2: Time series of reported dengue cases in Northern Mindanao from 2009 to 2010.

The estimated model is *Equation* 1 with $D = 8$ and $X_{t,d} = 0$ because no covariates information on variables was available. As illustrated in Figure 2 (green dashed line) and accompanied by 95% prediction intervals, the model was used to correct the total number of instances N_t not just for that week but also for the seven weeks prior. The graph demonstrates that the forecasts detect the fact that the number of dengue cases increased. In addition, it can be seen that the model forecast (green dotted line) and reported cases (dark blue line) increase in a similar fashion up to the 25th week of 2010.

Considering the present number of instances, a decision-maker in public health could take the wrong decisions in June 2010 because it looks to be declining. Nonetheless, it can be noticed that the eventually reported instances are corrected for delays and, on occasion, misclassification utilizing laboratory confirmation tests or other factors that may cause delayed reports.

B. Model Evaluation

A set of checks are conducted to ensure that the model accurately represents the data. These include the deviation of the prediction from the actual value, the examination of predictive samples of the total, the sample means and sample variance of the total N_t, and the temporal dependency of the data.

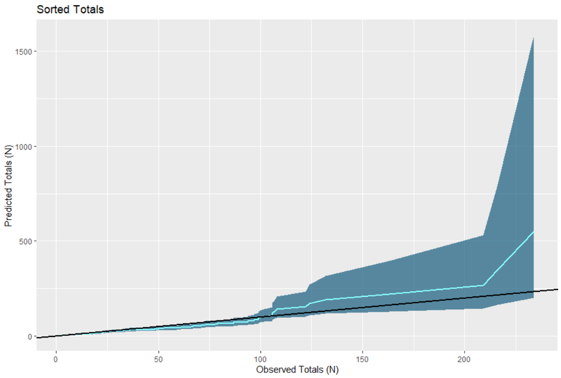


Fig. 3: Predicted totals plotted against the respective observed (sorted) values.

Figure 3 depicts the predicted N_t, defined as the means of these distributions, versus the observed N_t, sorted ascendingly. The black line represents the $x = y$ equation, and it shows how far the prediction deviated from the actual value. The blue line shows where the points would fall if all predicted values perfectly matched the observed ones, and 95% prediction intervals were also included. Figure 2 shows that the model estimates for dengue cases in Northern Mindanao very well capture the rank of the observed values, despite the fact that eight weeks of the 77 values are based on data that the model has not seen.

Second, to evaluate the accuracy of the prediction samples of the totals. Figure 4 depicts the predictive distributions for the sample mean and standard deviation of the totals N_t.

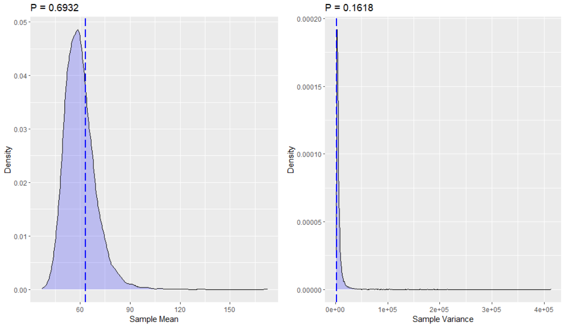


Fig. 4: Sample mean and sample variance

The observed values are represented by the vertical lines, while the tail portions of the observed values are represented by the probabilities that are quoted (values less than 0.025 or over 0.975 suggest that the model does not represent the observed value well). Figure 4 shows that the sample mean, and variance are accurately represented, given that the sample mean is 0.6932 and the sample variance is 0.1618. This implies that the sample mean, and variance are accurate (like, are not extreme concerning the distribution).

When it comes to being able to detect breakouts, temporal dependence is necessary. The degree of dependence that exists between different points in the time series is one of the factors that must be taken into consideration. Figure 5 illustrates the sample autocorrelation in the N_t for each of the eight lags, which may be used to determine whether or not the temporal dependency in N_t is accurately represented.

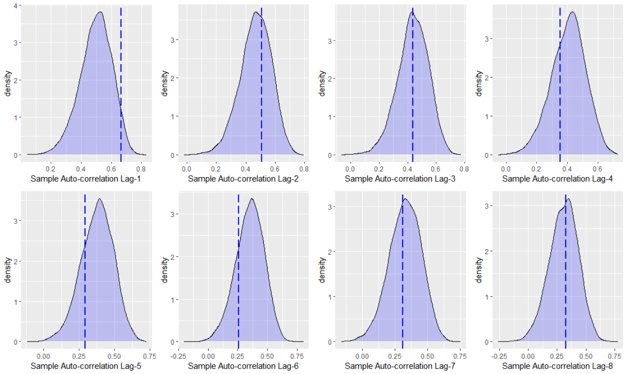


Fig. 5: Predictive distribution for the sample autocorrelation of the totals.

Figure 5 demonstrates that the model well reflects the time dependence in N_t, since none of the observed values (vertical lines) are extreme in relation to their respective prediction distributions. It is widely accepted that our model's estimate of dengue cases is significantly dependent on the recent trend of the number of cases. This suggests that it can be utilized for behavior prediction, as it offers an indication as to how the number of dengue cases would behave in the near future.

C. Estimate and Rolling Prediction

Figure 6 depicts the estimations of three variables, namely α_t the overall temporal evolution of the counts, β_d the delay structure, and $\eta_{w(t)}$ the seasonal variability.

$$log(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)} \tag{10}$$

Figure 6(a) indicates that the overall temporal effect grows initially by a little amount, then continues to rise gradually, possibly representing an increase in the vulnerable population. In Figure 6 (b), the delay structure initially increases, then diminishes; as would be predicted, as time passes, more cases are recorded. However, in Figure 2, the eventually reported dengue cases exhibit no seasonality; it is therefore not surprising that Figure 6 (c) captures no seasonal component.

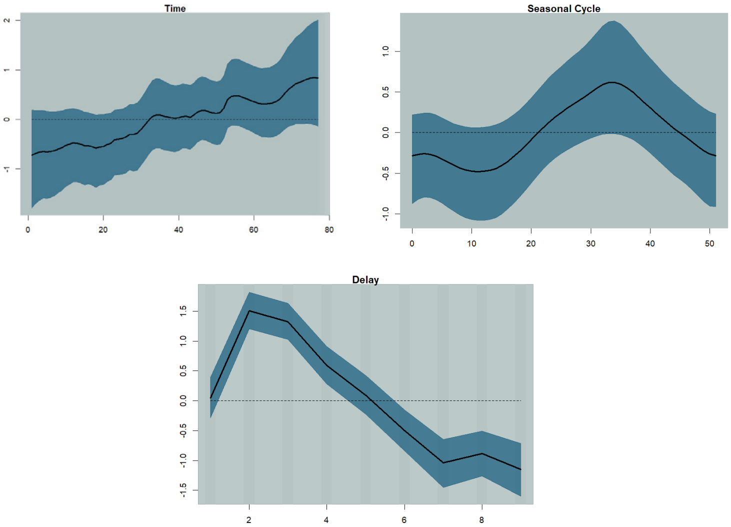


Fig. 6: Estimates of the overall temporal variation, overall delay structure and seasonal variability.

The weekly rolling forecasts from the 25th week of 2010 (June 14–20) to the 36th week of 2010 (August 16–22) are shown in the following figures. The black line represents the number of instances finally reported, the red dashed line represents the number of cases now reported, and the green dashed line represents the model prediction along with 95% prediction intervals. The symbol of a circling cross represents the weeks T=77, 78…,88. This period was chosen expressly to evaluate the model's capacity to capture an outbreak and the subsequent significant fall in the number of reported cases (black line).

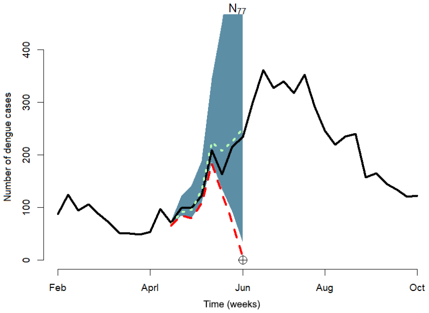


Fig. 7 N_{77} (77$^{\text{th}}$ week prediction)

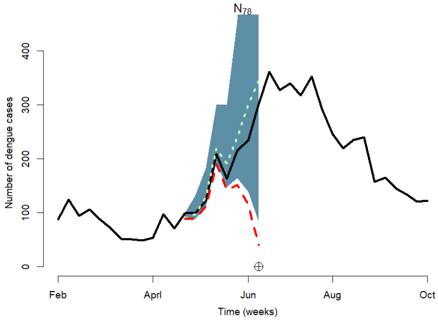


Fig. 8 N_{78} (78$^{\text{th}}$ week prediction)

Figures 7, 8, and 9 depict the time series for predicting dengue cases in Northern Mindanao from the 77th to the 78th week. The currently reported cases from the 25th to the 27th week of 2010 appear to be decreasing (red dashed line). Nonetheless, immediate calculations indicate that the number of dengue cases is rising (solid black line). Based on the cases reported, the outcome could lead to the wrong actions or decisions, making people think that no preventive measures are needed since the number of dengue cases is decreasing. On the other hand, the model (green dashed line) accurately captures the upward trend of eventually reported dengue cases. Thus, it is sufficient to demonstrate that the one-to-three-week-ahead forecast accurately predicts the number of dengue cases.

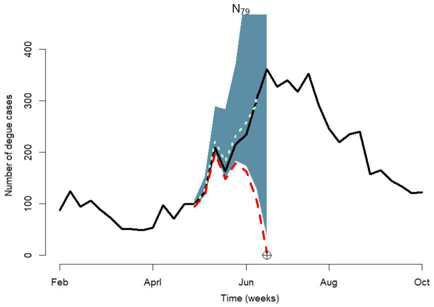


Fig. 9 N_{79} (79th week prediction)

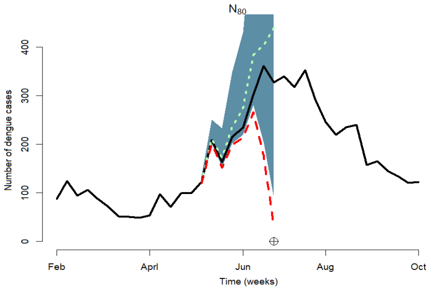


Fig. 10 N_{80} (80th week prediction)

Figures 10, 11 and 12 shows the time series for predicting dengue cases in Northern Mindanao from the 80th to the 82nd week. The outcome suggests that the number of cases reported during the 28th to 30th week of 2010 appears to be decreasing (red dashed line). Despite this, the calculation reveals that the number of dengue cases is significantly higher (solid black line) than the currently reported cases when there is no delay.

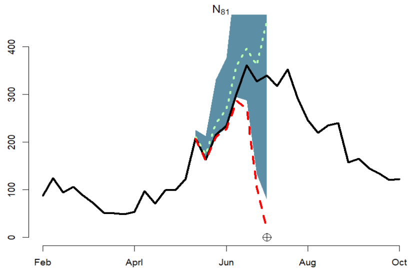


Fig. 11 N_{81} (81th week prediction).

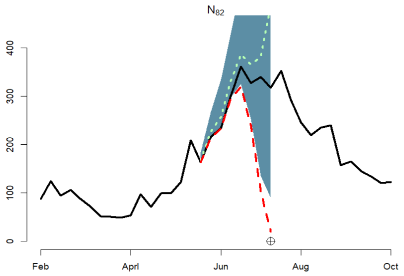


Fig. 12 N_{82} (82th week prediction)

In addition, the model (green dotted line) in Figures 10 and 12 predicts the increasing trend but misses a slight decrease in the eventually reported cases; the estimate predicted a greater number of dengue cases than the actual counts, particularly in Figure 11. However, the model prediction from N_{85} to N_{88} does not capture the eventually reported number of cases as it decreases its counts.

In summary, Figures 7, 8, and 9 demonstrate that in N_{77}, N_{78}, and N_{79} weeks, the model (green dotted line) captures the increase in the eventually reported number of cases. Figures 10, 11 and 12 demonstrate that the model projected a higher number of dengue cases in the years N_{80} to N_{82} than the actual eventually reported number of dengue cases.

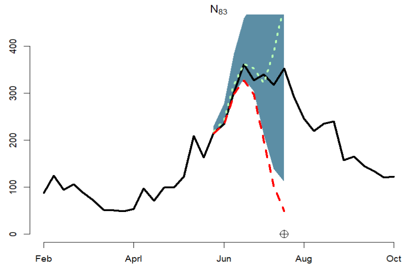


Fig. 13 N_{83} (83rd week prediction).

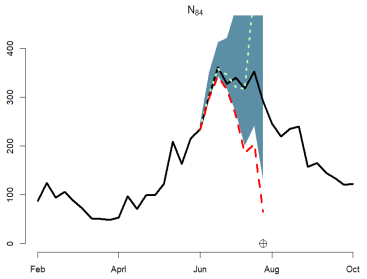


Fig. 14 N_{84} (84th week prediction)

Figures 13, 14, 15, 16, 17 and 18 demonstrate that in N_{83} to N_{88}, the model prediction of the number of dengue cases appears to increase, as the decreasing trend is not captured. Since then, most of the reported counts have fallen within the 95% prediction intervals, especially for time T (shown by the circled cross), which is the most important value.

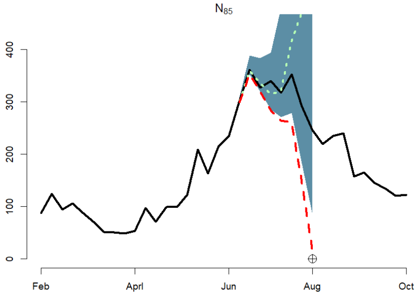


Fig. 15 N_{85} (85th week prediction)

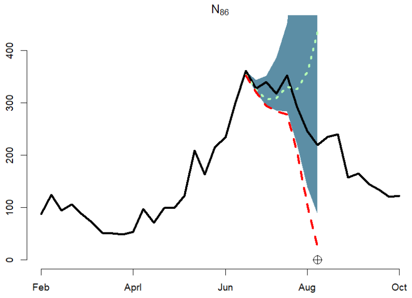


Fig. 16 N_{86} (86th week prediction)

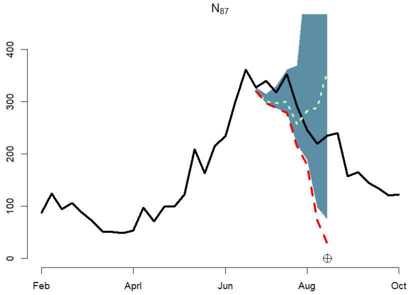


Fig. 17 N_{87} (87th week prediction)

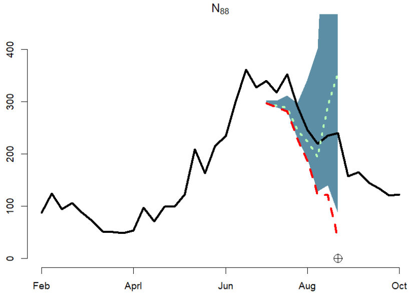


Fig. 18 N_{88} (88th week prediction)

V. CONCLUSION

The Bayesian Hierarchical Approach is a versatile paradigm for time-delayed structured illness count data applications. This enables the estimation of the missing (observable) data as $n_{t,d} \sim NegBin(\lambda_{t,d}, \phi)$ so that nowcasting can be performed. Dengue data from Northern Mindanao are utilized to examine and illustrate the performance of the model, demonstrating the required flexibility and complexity of the framework. Temporal dependence is essential for detecting epidemics; despite the absence of covariates in the model, the R-INLA implementation makes this a simple operation. Because we employ the Laplace approximation (INLA) to obtain samples from the (marginal) posteriors, implementing the models in the Bayesian framework is exceedingly rapid.

Surveillance and warning systems that rely on reported incidence to gauge danger may be uninformed if the delay is not addressed; hence, accurate estimates are crucial. Consequently, this strategy can be a useful tool for making decisions in surveillance systems. Forecasting based on the predictive distribution of counts suggests that the presented models' predictions can be readily incorporated into a decision-theoretic framework for issuing alerts.

LITERATURE CITED

BASTOS LS., ECONOMOU T., GOMES M., VILLELA D., COELHO F., CRUZ O., STONER O., BAILEY T., and CODECO C., 2019, A Modelling Approach for Correcting Reporting Delays in Disease Surveillance Data. Statistics in Medicine. 38:4363–4377.

BROOKMEYER R., and DAMIANO A., 1989, Statistical Methods for Short-Term Projections of AIDS Incidence, Stat Med;8(1):23–34

CODECO C. T., COELHO F., CRUZ O., BASTOS LS., OLIVEIRA S., and CASTRO T., 2019. Infodengue: A Nowcasting System for the Surveillance of Arboviruses in Brazil. Revue d'Épidémiologie et de Santé Publique, V. 66, https://doi.org/10.1016/j.respe.2018.05.408

FARRINGTON C.P., ANDREWS N.J., BEALE A.D., and CATCHPOLE M.A., 1996, A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. J Royal Stat Soc, Ser (Stat Soc). 1996;159(3):547-563.

HOHLE M., and HEIDEN M., 2011, Bayesian Nowcasting During the STEC O104:H4 Outbreak in Germany. Biometrics. 70(4):993,1002.

KLAUCKE DN., BUEHLER JW., THACKER SB., PARRISH G., TROWBRIDGE FL., and BERKELMAN RL., 1988, Guidelines for Evaluating Surveillance Systems. Morb Mortal Wkly Rep. 37(Suppl5):1-18.

LAWLESS J.F., 1994, Adjustments for Reporting Delays and the Prediction of Occurred But Not Reported Events. The Canadian Journal of Statistics, 22(1), 15–31.

LIN H., YIP PS., and HUGGINS R., 2008, A Double-Nonparametric Procedure for Estimating the Number of Delay-Reported Cases. Stat Med.;27(17):3325–39.

MACK T., 1993, Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates. ASTIN Bulletin. 23(2):213-225.

MIDTHUNE DN, 2005; Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data. Taylor & Francis, Ltd.;100(469):61–70.

NOUFAILY A., FARRINGTON P., GARTHWAITE P., ENKI DG., ANDREWS N., and CHARLETTE A., 2016, Detection of Infectious Disease Outbreaks from Laboratory Data with Reporting Delays. J Am Stat Assoc.;111(514):488–99.

NOUFAILY A., ENKI D., FARRIGTON P., GARTHWAITE P., ANDREWS N., and CHARLETT A., 2013, An Improved Algorithm for Outbreak Detection in Multiple Surveillance Systems. Stat Med.;32(7):1206–22.

RENSHAW AE., and VERRAL R.J. 1998, A Stochastic Model Underlying the Chain-Ladder Technique. British Actuarial Journal; 4(04):903-923.

ROSINSKA M., PANTAZIS N., JANIEC J., PHARRIS A., AMATO-GAUCI AJ., and QUINTEN C., 2018, Potential Adjustment Methodology for Missing Data and Reporting Delay in the HIV Surveillance System, European Union/European Economic Area.

SALMON M., SCHUMACHER D., STARK K., and HOHLE M., 2015, Bayesian Outbreak Detection in The Presence of Reporting Delays. Biometrical Journal. 57(6):1051-1067.

UNDURRAGA E., EDILLO F., ERASMO J.V.E., ALERA M.T., YOON I.K., LARGO F.M., and SHEPARD D.S., 2017. Disease Burden of Dengue in the Philippines: Adjusting for Underreporting by Comparing Active and Passive Dengue Surveillance in Punta Princesa, Cebu City. The American Society of Tropical Medicine and Hygiene.