Bayes Count Data Forecasting with Rainfall as Covariate for Dengue Fever Cases in South Sulawesi

Asrirawan¹ and Khaerati²

¹Mathematics Department, Natural Science Faculty, Cokroaminoto Palopo University, Indonesia.
²Biology Education Department, Education Faculty, Cokroaminoto Palopo University, Indonesia

E-mail : ¹enalmantovani@gmail.com & Khaerati89@gmail.com

Abstract. An accurate early warning system to forecast dengue fever enhances the effectiveness of preventive measures against dengue fever and the most widely used is time series analysis. Forecasting dengue by using bayes count data is a need for improved surveillance tools. In this paper we constructed a count data model with including rainfall as covariate. Rainfall slightly improved dengue fever forecasting in South Sulawesi by bayes count data model which was GSARIMA-X fitted to logarithmically transformed dengue fever case data twelve ahead. In addition, we compare accurately the result of forecasting either with rainfall as covariate or not.

1. Introduction
In recent years, cases of Dengue Hemorrhagic Fever (DHF) usually appear in the transition season, especially in the early months of the year. In 2014, until mid-December, it recorded dengue cases in 34 provinces in Indonesia 71,668 people, and 641 of them died. The number one was lower than the previous year, namely in 2013 with the number of people as many as 112,511 people and the number of cases died 871 patients.However, Dengue Hemorrhagic Fever is still one of the main health problems faced by Indonesia [1].

One solution to reduce the number of DHF patients makes early prevention by doing forecasting as early as possible so certain moments can be known the months of any case of dengue fever up or down. Count data forecasting with Gaussian approach (normal) has been done by Directorate of Anti Malaria Campaign, Health Minister in Sri Lanka. One of all models that can be used is Generalized Seasonal Autoregressive Integrated Moving Average (GSARIMA). It relates between ARMA component models and predictor variable to mean parameter transformation of the data distribution by link function. It is used to show that the data distribution remains in the domain of positive real numbers, so it can get more accurate prediction. However, it can only be applied to stationary and non-seasonal time series data. Benjamin [2] used approaches of iteratively reweighted least squares (IRLS) to estimate the parameters of GARMA models.

Briet, Amerasinghe, and Vounatsou [3] developed the generalized model of seasonal autoregressive integrated moving average (GSARIMA) models involving seasonal effects and
differencing orders and applied to malaria time series data in Sri Lanka. Briet caught up rainfall as predictor variable and assumed the malaria time series data following the negative binomial distribution. In this research, the GSARIMA models will be estimated using Bayesian approach and compared the accurate rate of forecasting models with SARIMA models. It was applied to the number of dengue hemorrhagic fever (DHF) patients in South Sulawesi. Asrirawan and Suhartono [4] forecasted the number of dengue fever patients in Surabaya using Bayesian estimation GSARIMA models. In addition, Iriani [5] said that there was a correlation between rainfall and an increase in the number of dengue cases. This weather change will cause modifications in the Aedes aegypti mosquito habitat. A prediction model of the number of DHF events should be developed to anticipate or determine what actions the government should take as efforts to overcome the incidence of DHF.

2. Method

2.1. Case Data
GSARIMAX models will be applied to the number of dengue hemorrhagic fever (DHF) patients. The data are obtained at the Public Health Office, South Sulawesi and Meteorology, Climatology, and Geophysics Office. The amount of data used 168 counts from January 2003 to December 2016. The Bayesian programs used to estimation real data are R (“JAGS”), and the classical ARIMA models use Minitab and SAS.

2.2. GSARIMAX Models
The GSARIMAX models are developed based on the GARMA models involving seasonal effects and differencing. Let \( y^T = (y_1, y_{t+1}, ..., y_{t+n}) \) be a time series of count data. It follows

\[
y_t \sim \text{NegBin}(\mu_t, \psi) \quad \text{with} \quad E(y_t) = \mu_t \quad \text{and} \quad \text{Var}(y_t) = \mu_t + \frac{\mu_t^2}{\psi}.
\]

The GARMA models can be seen:

\[
g(\mu_t) = \phi_p(B)X_t^T \beta - g(y_t) + g(y_t) - g(\mu_t)
\]

\[
\text{where} \quad g(\mu_t) \text{is the link function,} \quad \phi_p(B) = 1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p \quad \text{and} \quad \theta_q(B) = 1 - \Theta_1 B - \Theta_2 B^2 - ... - \Theta_q B^q, \quad B \text{ is the backshift operator with } B^d y_t = y_{t-d} \text{. The } \beta^T = (\beta_0, \beta_1, \beta_2, ..., \beta_s) \text{ vector is the coefficient for the } X_t^T = \left(x_0, x_{1,t}, x_{2,t}, ..., x_{s,t}\right) \text{ vector with } x_0 \text{ as an intercept with } x_0 = 1 \text{. In GARMA models, the count data modeled using a logarithmic link function or identity [6]. If } y_t \text{ follows a Poisson distribution with a mean value as follows:}
\]

\[
\mu_t = E(y_t | y_{t-1}) = \mu_0 \exp(- \beta_0 - \beta_1 y_{t-1}), \quad \beta_i > 0
\]

then there are two methods of transformation that can be used as follows:

1) ZQ1 link function transformation has the form:

\[
\log(\mu_t) = X_t^T \beta + \sum_{i=1}^{q} \theta_i \left[\log(y_{t-i}^\tau) - X_{t-i}^T \beta\right]
\]

\[
\text{with} \quad y_{t-i}^\tau = \max(y_{t-i}, c), 0 < c \leq 1.
\]
2) ZQ2 link function transformation has the form:

\[
\log(\mu_t) = X_t^\top \beta + \sum_{i=1}^{q} \theta_i \left[ \log(y_{t-i} + c) - \log[\exp(X_{t-i}^\top \beta) + c] \right]
\]

The negative binomial ZQ1 GSARIMA models be given:

\[
\log(\mu_t) = \phi_p(B)(1 - B)^p(1 - B^s)^q \Phi_p(B^s)[X_t^\top \beta - \log(y_t^\top)] + \log(y_t^\top) - \theta_q(B)\Theta_q(B^s)\times
\]

\[
\log\left( \frac{y_t^\top}{\mu_t} \right) + \log\left( \frac{y_t^\top}{\mu_t} \right)
\]

and ZQ2 as follows:

\[
\log(\mu_t) = \phi_p(B)(1 - B)^p(1 - B^s)^q \Phi_p(B^s)[\log[\exp(X_t^\top \beta) + c] - \log(y_t + c)] + \log(y_t + c) - \theta_q(B)\Theta_q(B^s)\log\left( \frac{y_t + c}{\mu_t + c} \right) + \log\left( \frac{y_t + c}{\mu_t + c} \right)
\]

3. Results and Discussions

3.1. Model Fit

In this paper, GSARIMA models are estimated in bayesian framework based on Briet. In bayesian inference, all model parameters are changed for making prior distributions. A weakly stationary model is assumed and, therefore, the autocorrelation and moving average parameters are constrained using an algorithm. For this purpose, the autoregressive and moving average parameters in the likelihood were reparameterized and prior distributions were adopted on the new parameterization. For example, the non seasonal autoregressive parameters \( \phi_1, \phi_2, ..., \phi_p \) are reparameterized in terms of \( r \),

\[
r_t = \begin{pmatrix} r_1, r_2, ..., r_p \end{pmatrix}, \text{ where } \phi_p = 2r_1 - 1 \text{ and } \phi_{p-i} = 2r_{p-i} - 1 - \sum_{k=1}^{i} (2r_{p-k} - 1)(2r_{p-k+1} - 1). \]

The following prior distributions are assumed

\[
r_i \sim Beta\left[ \frac{1}{2} (i + 1), \frac{1}{2} (i + 1) \right], i = 1, 2, K, p
\]

Further priors chosen are \( \beta_0, \beta_1, K, \beta_p \sim N(0,1000) \) and \( \psi \sim Ga(0.01,0.01) \). In bayesian context, the joint posterior distribution with the first observation products between likelihood and prior distribution by using parameters. Given

\[
\phi, \varphi, 0, \varphi^*, 0^* | y_1, ..., y_n, X, \psi, \beta, \alpha, \theta^* \theta^*
\]

\[
P(y_{w+1}, ..., y_n | y_1, ..., y_n, X, \psi, \beta, \varphi, 0, \varphi^*, 0^*)p(\psi)p(\beta)p(\alpha)p(\theta|\theta^*)p(\theta^*)
\]

Within X is nxv+1, \( \varphi^T = (\varphi_1, ..., \varphi_p) \), \( \varphi^T = (\varphi_1, ..., \varphi_q) \) and \( \theta^T = (\theta_1, ..., \theta_q) \).

The GSARIMAX models were estimated using the free Bayesian software programme, “JAGS”, which employs Markov chain Monte Carlo (MCMC) simulation methods. Examples of code written for using JAGS within the R software, for negative binomial GSARIMAX models with logarithmic
The link function and ZQ1 transformation, are provided as supporting information [4]. For the evaluation of model, it can be used to MARE criteria that have the form:

$$MARE = \frac{1}{N-1} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i + 1}$$

### 3.2. Application to Dengue Hemorrhagic Fever Analysis

In this paper, this section displays forecasting DHF count data time series using GSARIMAX. To obtain this model has to be initially performed to SARIMA model for determining the seasonal and non-seasonal parameters. The initial step is to identify the order of SARIMA \((p, d, q)(P, D, Q)^\Delta\). Based on the results of Box-Cox transformation, it turns out of the data that has not been stationary in variance with the rounded value of 0.25. After the transformation, it can be seen that the data have shown stationary in variance.

Figure 1 shows that the data have not been stationary at lag of non-seasonal and seasonal, so we have to difference data. After differencing the lag 1 and 12, it can be seen in figure 2 that the data have shown a stationary mean both seasonal and non-seasonal lags.

![Figure 1. Time Series Plot and Autocorrelation Function (ACF)](image1)

![Figure 2. ACF and PACF Plots after differencing seasonal and non-seasonal lags](image2)
The next step is the estimation and significance tests for ARIMA model parameters begun by looking at the ACF and PACF plots in figure 2. ACF plot indicates the lag significant only 1 and 12 while PACF plot shows lag significant only in 1 and dies down on seasonal lags (12, 24, 36 and 48). The models after identification that can be formed is a subset SARIMA models. After white noise test, it does not turn out to meet the assumptions of the residual white noise model so that the necessary additional lag was in the model. As the results of model preminilary identification, the DHF data follow SARIMA (1,1,1)(1,1,0)\textsuperscript{12}. SARIMA (1,1,1)(1,1,1)\textsuperscript{12} and SARIMA (1,1,1)(0,1,1)\textsuperscript{12} models.

The results of parameter estimation and significance tests with p-value of the parameter are less than 5% error level. SARIMA values that have the smallest AIC values have chosen as the best model.

Based on Table 1, it can be seen that GSARIMA-X (1,1,1)(1,1,0)\textsuperscript{12} model with rainfall as covariate is the best model than these models. The next step is the estimation of DHF count data time series using MCMC algorithms implemented in JAQS software. After analysis, it can be seen in Table 2.

For remaining the best link function, the models are evaluated based on DIC criteria. The table provides that ZQ1 transformation is relatively better than ZQ2 transformation. The ZQ1 GSARIMA model can be written as follows:

\[
\log(\mu_t) = (1 - \phi_1 B)(1 - B)(1 - B^{12})(1 - \phi_2 B^{12})(1 - \beta_t X) - \log\left(\frac{\bar{y}_t}{\mu_t}\right) + \log\left(\frac{\bar{y}_t}{\mu_t}\right) - (1 - \theta_1 B)\times \\
\log\left(\frac{\bar{y}_t}{\mu_t}\right) + \log\left(\frac{\bar{y}_t}{\mu_t}\right)
\]

Model GSARIMAX (1,1,1)(1,1,0)\textsuperscript{12} transformasi ZQ2

\[
\log(\mu_t) = (1 - \phi_1 B)(1 - B)(1 - B^{12})(1 - \phi_2 B^{12})\log[\exp(1 - \beta_t X) + c] - \log(y_t + c)] + \log(y_t + c) - \\
(1 - \theta_1 B)\log\left(\frac{\bar{y}_t + c}{\mu_t + c}\right) + \log\left(\frac{\bar{y}_t + c}{\mu_t + c}\right)
\]

The 4th International Seminar on Sciences  IOP Publishing
IOP Conf. Series: Earth and Environmental Science 187 (2018) 012047  doi:10.1088/1755-1315/187/1/012047

| Parameter | GSARIMA-X (1,1,1)(1,1,0)\textsuperscript{12} |
|-----------|-------------------------------------|
| $\beta$   | -0.345 (-0.678;0.025)                |
| $\phi_1$  | -0.214 (-0.389;0.045)                |
| $\theta_1$| 0.156 (0.129;0.238)                  |
| $\phi_2$  | 0.417(0.324;0.526)                   |
| $\psi$    | 4.315(3.752;5.673)                   |

Table 1 Criteria for Selection of the best SARIMA models

| Model                  | DIC    | Maref  |
|------------------------|--------|--------|
| GSARIMA-X (1,1,1)(1,1,0)\textsuperscript{12} | 3650.70 | 0.3291 |
| GSARIMA-X (1,1,1)(1,1,1)\textsuperscript{12} | 3656.82 | 0.3386 |
| GSARIMA-X (1,1,1)(0,1,1)\textsuperscript{12} | 3670.76 | 0.3421 |

Table 2 Parameter estimation (mean and 95% credible interval) of selected negative binomial models.
The authors acknowledge the Public Health Office in South Sulawesi for making surveillance data available and Ministry of Research, Technology and Higher Education of the Republic of Indonesia contributed to gave the cost for researcher.

References
[1] Ministry of Health of Indonesia 2015 Epidemiology Bulletin Volume 2 Jakarta
[2] Benjamin, M. A., Rigby R. A., & Stasinopoulos, D. M. 1998 Fitting Non-Gaussian Time Series Models COMPSTAT Proceedings in Computational Statistics, eds. R. Payne dan P.Green, Heldelburg: Physica-Verlag 191-196
[3] Briet, J. T. O., Amerasinghe, H. P., & Vounatsou, P. 2013 Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Number Malaria Journal PLoS ONE 8(6):e65751
[4] Asrirawan & Suhartono 2015 The estimation of Generalized Seasonal Autoregressive Integrated Moving Average (GSARIMA) Models using by Bayesian Approach for Forecasting The Number of DHF Patients. Proceeding ICMSTR
[5] Iriani, Yuli 2012 Relationship between Rainfall and Increase Case of Dengue Hemorrhagic
Fever Children in Palembang City Journal volume 13 Palembang: Faculty of Medicine Universitas Sriwijaya

[6] Zeger, S. L., & Qaqish, B., 1988 Markov Regression Models for Time Series: A Quasi-Likelihood Approach. Biometrics 44 1019-1031