Forecasting dengue hemorrhagic fever cases using ARIMA model: a case study in Asahan district

Fazidah A Siregar¹, Tri Makmur², Saprin S³
¹Faculty of Public Health, University of Sumatera Utara, Medan, 20155, North Sumatera, Indonesia
²Faculty of Medicine, Islamic University of Sumatera Utara, Medan, North Sumatera, Indonesia
³Asahan district health department, North Sumatera Province

*Email: fazidah@usu.ac.id

Abstract. Time series analysis had been increasingly used to forecast the number of dengue hemorrhagic fever in many studies. Since no vaccine exist and poor public health infrastructure, predicting the occurrence of dengue hemorrhagic fever (DHF) is crucial. This study was conducted to determine trend and forecasting the occurrence of DHF in Asahan district, North Sumatera Province. Monthly reported dengue cases for the years 2012-2016 were obtained from the district health offices. A time series analysis was conducted by Autoressive integrated moving average (ARIMA) modeling to forecast the occurrence of DHF. The results demonstrated that the reported DHF cases showed a seasonal variation. The SARIMA (1,0,0)(0,1,1)₂ model was the best model and adequate for the data. The SARIMA model for DHF is necessary and could applied to predict the incidence of DHF in Asahan district and assist with design public health measures to prevent and control the diseases.

1. Introduction
Dengue hemorrhagic fever (DHF) still health problem worldwide and currently more than 100 countries are endemic for dengue virus infection with an estimated 50 million dengue infections occur annually [1,2]. According to World Health Organization (WHO), Indonesia ranked second for DHF cases in Asia. The Ministry of Health of Indonesia reported that North Sumatera province ranked third highest for DHF cases with 16 dengue endemic districts [3]. Asahan district is one of them with the incidence of DHF increased annually, from 31/100,000 population in 2012 to 102/100,000 population in 2016 [4].

The DHF is a complex problem. The dynamic of dengue infection are driven by complex interaction between host, vectors and viruses that are influenced by environmental. On the other hand, many factors contribute to DHF incidence, including socio-demographic, behavioural, cultural and environmental factors. Moreover, Aedes mosquito as the main vector for DHF play role on transmission also influenced by climate and environmental [5,6].

Currently, no specific treatment and no vaccines are available for dengue hemorrhagic fever. In addition, lack of effective vector control measures and poor disease surveillance, made the dengue prevention measures may often be implemented late, thereby reducing the opportunities for preventing transmission and controlling the epidemic. DHF outbreaks can be predicted by epidemiological modeling thus enabling the health systems to be prepare to manage outbreaks. Therefore, enhancing the preparedness by predicting the occurrence of DHF made crucial. Time series methodology as forecasting model has been increasingly used in the field of epidemiological research of infectious disease, such as malaria [7,8] dan dengue hemorrhagic fever [9-13].

Detailed information about when and where DHF occurred in the past can be useful to describe the magnitude problem and predict the disease in the future time. This could assist public health intervention to prevent and control the disease. This study was undertaken to develop a prediction model for DHF using time series data over the past decade in Asahan district and to forecast the monthly DHF incidence for the year 2017.
2. Methods
This study was conducted in Asahan district, an endemic area for DHF cases in North Sumatera Province. Asahan is situated south of North Sumatera Province, situated between 2.300 – 3.260 North Latitude and 99.01- 100.000. It covers a surface area of 4,580,75 km2 with total population (2010) was 1,024,369 (Figure 1).

2.1. Data collection
Reported monthly dengue hemorrhagic cases for January 2012 through December 2016 were obtained from Asahan district health offices.

2.2. Statistical analysis
Autoregressive integrated moving average (ARIMA) models have been used for analyzing time series data containing seasonal trends to develop a forecasting model that was first popularized by Box and Jenkins [14]. This method for selecting an appropriate ARIMA model for estimating and forecasting an univariate time-series consisted of identification, estimation, diagnostic checking and forecasting. ARIMA model may include autoregressive (p) terms, differencing (d) terms and moving average (q) operations and is represented by ARIMA (p, d, q). A stationary time series is one whose statistical properties such as mean and variance are constant over time. Seasonality usually causes the series to be nonstationary. Therefore, Seasonal ARIMA (SARIMA ) is an extension of the ARIMA method to a series in which a pattern repeats seasonally over time and is represented as SARIMA (p, d, q) (P, D, Q)s, which that the Seasonal autoregressive (P), seasonal differencing (D), and seasonal moving average parameters (Q); s defines the number of time periods until the pattern repeats again for a monthly data it is.

Minitab version 16.0 was used to analysis and determine the best-fitting model. First, check for stationary was made with the aid of a control chart. After verifying that the series was stationary, an ARIMA model was developed. Before using the model for forecasting, it was checked for adequacy. This was done through examining the ACF and PACF of the residuals. Further, Ljung-Box test was used to provide an indication of whether the model was correctly specified. A significant value less than 0.05 was considered to acknowledge the presence of structure in the observed series, which was not accounted for by the model; therefore, we ignored the model if it had significant value. After the best model was identified, forecast for monthly values of the year 2017 were made.

3. Results and Discussions
The reported monthly DHF cases during 2012-2016 exhibited a seasonal pattern (Figure 2) and increased significantly during the study period (Figure 3). The number of monthly DHF cases began to increase in June, peaked in October and then gradually decreased.

![Figure 1. Asahan district in North Sumatera Province](image-url)
Figure 2. Monthly number of DHF cases in Asahan district from 2012 to 2016

Figure 3. Trend analysis of the number of DHF cases in Asahan district from 2012 to 2016

The time series plot of the reported DHF cases displayed seasonal fluctuations and therefore deemed non stationary. The ACF of DHF time-series and the PACF of DHF time series are expressed in Figure 4. Decrease in autocorrelation values after lag 2 indicated non stationary data; consequently, there was need to include a first-lag difference term in the SARIMA model structure (d = 1) as presented in Figure 5.

Figure 4. ACF and PACF of the reported DHF cases
Among the statistical models, SARIMA (1,0,0) (0, 1, 1) was selected as the best model, with the lowest mean square error of 442 and p value (L Jung-Box) 0.602. Ljung-Box test > 0.05 suggested that there were no significant autocorrelation between residuals at different lag times and the residuals were white noise [Table 1]. This was further corroborated by plotting the ACF and PACF of the residuals as presented in Figure 6.

### Table 1. Characteristics of SARIMA models

| No | Model | p value | MS | p value |
|----|-------|---------|----|---------|
|    |       | AR      | SAR | MA | SMA | (LJung Box) |
| 1  | (1,0,0)(1,0,0): | 0.000 | 0.684 | - | - | 424 | 0.451 |
| 2  | (1,0,0)(0,1,1): | 0.000 | - | - | 0.013 | 442 | 0.602 |
| 3  | (0,1,1)(0,1,1): | - | - | 0.326 | 0.005 | 478 | 0.575 |
| 4  | (1,1,0)(0,1,1): | 0.756 | - | - | 0.010 | 507 | 0.779 |

The SARIMA model was constructed using data collected between January 2012 until December 2016 and was verified using data collected from January to June 2017. The validity of model was tested and used to forecast monthly DHF cases in 2017. The result show that the actual value and the predicted value matched reasonably well as presented in Figure 7.

**Figure 5.** ACF and PACF of DHF difference first lag

**Figure 6.** ACF and PACF of the residual of SARIMA (1,0,0)(0,1,1)
Figure 7. The actual monthly DHF cases from 2012 - 2016 and the monthly DHF cases predicted for 2017 by using SARIMA model in Asahan district

ARIMA models useful in modelling dependence of a time series between observation [15] and to be adequate tools for use in epidemiological surveillance [16]. Our study provides an example of applying a SARIMA model to forecast incidence of DHF in Asahan district. These models have been utilized to forecast DHF incidence in several countries [9,10,12,16,17].

Among all models, SARIMA (1,0,0) (0,1,1)$_{12}$ was the most suitable predictive model in this study, which showed the lowest mean square error. A study in Brazil found SARIMA (2,1,3)(1,1,1)$_{12}$ model was best fit for the dengue incidence data [10]. While a study by Luz et al in Rio de Janeiro, Brazil found that SARIMA (2,0,0) (1,0,0)$_{12}$ model was deemed best fit and no seasonal differencing [9]. Choudhury et al., reported SARIMA (1,0,0) (1, 1, 1)$_{12}$ as the most suitable model for forecasting dengue incidence in Dhaka, Bangladesh [18]. On another study in Thailand found SARIMA (2,0,1) (0, 2, 0)$_{12}$ , ARIMA (1,0,1), and SARIMA (2, 1, 0) (0, 1, 1)$_{12}$ models as most suitable for DHF cases [17,19,20].

ARIMA modeling is a useful tool for interpreting surveillance data and forecast of the cases to help guide timely prevention and control measures. In addition, may providing decision-makers a clear idea in designing an effective prevention and control strategy for DHF. Further research is recommended to integrate the forecasting model into the existing disease control program in terms in reducing the disease occurrence.

4. Conclusions
SARIMA (1,0,0)(0,1,1)$_{12}$ model provides the predicted incidence of DHF in Asahan district. Application this model could be used to predict the incidence of DHF in the future time and assist with design public health measures to prevent and control the disease.

Acknowledgments
We are grateful to the head of the Asahan district health office and staff for their assistance in this study. We are also thankful to all of the participants and people who were involved in this study.

References
[1]. Guzman, A. and Istüriz, R. E. (2010. Update on the global spread of dengue. International journal of antimicrobial agents 36 S40
[2]. Guzman, M. G et al (2011). Dengue: a continuing global threat. Nat Rev Microbiol. 8 S7
[3]. Ministry of Health of Indonesia. (2011). Report of dengue hemorrhagic cases in Indonesia, Jakarta, Indonesia.

[4]. Asahan district Health Department. (2016). Report of dengue hemorrhagic cases in Asahan district.

[5]. Rodriguez, H. & de la Hoz R. F. (2005). Dengue and vector behaviour in Ca’queza, Colombia, 2004. Revista de Salud Publica. 7(1) 1-15.

[6]. Halstead, S.B.(2008). Dengue virus-mosquito interactions. Annual Review Entomology. 53 273-291.

[7]. Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, and Kaewkungwal J. (2010). Development of temporal modelling forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. Malar J. 9 251.

[8]. Loha E, Lindtjorn B.(2010).Model variations in predicting incidence of Plasmodium falciparum malaria using 1998-2007 morbidity and meteorological data from south Ethiopia. Malar J. 9 166.

[9]. Luz, P. M., Mendes, B. V., Codeço, C. T., Struchiner, C. J and Galvani, A. P. (2008). Time series analysis of dengue incidence in Rio de Janeiro, Brazil. The American Journal of Tropical Medicine and Hygiene. 79(6) 933-939.

[10]. Martinez, E., Soares da Silva, E. & Dal Fabbro, A. (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. Revista da Sociedade Brasileira de Medicina Tropical 44(4) 436-440.

[11]. Lu, L., Lin, H., Tian, L., Yang, W., Sun, J. Liu.Q (2009). Time series analysis of dengue fever and weather in Guangzhou,ChinaBioMed Central Public Health 9(395) 1-5.

[12]. S. Wongkoon, M. Pollar, M. Jaroensutasinee, and K. Jaroensutasinee (2007). Predicting DHF Incidence in Northern Thailand using Time Series Analysis Technique International Scholarly and Scientific Research & Innovation 1(8) 484-488.

[13]. Bhatnagar S, Lai V, Gupta SD, Gupta OP. (2012). Forecasting incidence of dengue in Rajasthan, using time series analyses. Indian Journal of Public Health. 56(4) 281-285.

[14]. Box GE, Jenkins GM.Time series analysis: Forecasting and control. San Francisco:Holden-Day; 1976

[15]. Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., Ruche, G.L., Girdary, L. and Maramma, L. (2011). Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. BioMed Central Infectious Diseases,11(1) 1-13

[16]. Nobre FF, Monteiro AB, Telles PR, Williamson GD. Dynamic linear model and SARIMA(2001): A comparison of their forecasting performance in epidemiology. Stat Med 20 3051-69.

[17]. Promprou S, Jaroensutasinee M, Jaroensutasinee K. Forecasting dengue hemorrhagic fever cases in Southern Thailand using ARIMA models.(2006). Dengue Bull 30 99-106.

[18]. Choudhury MA, Banu S, Islam MA. Forecasting dengue incidence in Dhaka, Bangladesh (2008) : A time series analysis. Dengue Bull 32 29-37.

[19]. Silawan T, Singhasivanon P, Kaewkungwal J, Nimmanitya S, Suwonkerd W.(2008) Temporal patterns and forecast of dengue infection in Northeastern Thailand. Southeast Asian J Trop Med Pub Health 39 90-8.

[20]. Wongkoon S, Pollar M, Jaroensutasinee M, Jaroensutasinee K. (2008). Predicting DHF incidence in Northern Thailand using time series analysis technique. International Journal of Biological and Life Sciences 4 117-121.