Comparison of INAR(1)-Poisson model and Markov prediction model in forecasting the number of DHF patients in west java Indonesia

Atina Ahdika and Novyan Lusiyana

1Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Jalan Kaliurang Km 14.5 Sleman, Yogyakarta, Indonesia 55584
2Department of Parasitology, Faculty of Medicine, Universitas Islam Indonesia, Jalan Kaliurang Km 14.5 Sleman, Yogyakarta, Indonesia 55584

etina.a@uui.ac.id

Abstract. World Health Organization (WHO) noted Indonesia as the country with the highest dengue (DHF) cases in Southeast Asia. There are no vaccine and specific treatment for DHF. One of the efforts which can be done by both government and resident is doing a prevention action. In statistics, there are some methods to predict the number of DHF cases to be used as the reference to prevent the DHF cases. In this paper, a discrete time series model, INAR(1)-Poisson model in specific, and Markov prediction model are used to predict the number of DHF patients in West Java Indonesia. The result shows that MPM is the best model since it has the smallest value of MAE (mean absolute error) and MAPE (mean absolute percentage error).

1. Introduction
Dengue hemorrhagic fever (DHF) is one of the major health problems in Indonesia. DHF is caused by dengue virus, a positive-strand RNA virus of flaviviridae family with four distinct serotypes (DEN1-4) that transmitted by Aedes aegypti (Ae. aegypti) and Aedes albopictus (Ae. albopictus) [1]. DHF were also endemic, occurred throughout the year and can be accompanied by extraordinary events [2]. World Health Organization (WHO) states that the prevention and reduction of dengue virus transmission was strongly influenced by control vector and control human contact with mosquitoes, because until now there were not found a vaccine and specific treatment for DHF [3-4]. Although the treatment for DHF is not found yet, the government should know the estimation of the next cases which possibly happened to do a prevention action. The number of reported cases at time $t$ are usually influenced by the number of cases at time $t-1$. This kind of data can be predicted using time series model. Many researchers have estimated the number of DHF incidence using time series analysis. Promprou et al. (2006) used autoregressive integrated moving average (ARIMA) to model and forecast the monthly number of DHF cases in southern Thailand [5]. Choudhury et al. (2008) used seasonal autoregressive moving average (SARIMA) $(1,0,0)(1,1,1)_12$ to model the monthly number of DF cases in Dhaka, Bangladesh [6]. Martinez et al. (2011) used SARIMA $(2,1,2)(1,1,1)_12$ model to predict dengue incidence in Campinas, Southeast Brazil [7].Sriwattanapongse and Khanabsakdi (2011) used additive and multiplicative regression models to describe the patterns of hospital-diagnosed Malaria and DHF incidences by using the previous monthly or quarterly periods of incidences occurring in the upper Northern region of Thailand [8].
Based on the previous researches, the purpose of this paper is to predict the number of DHF cases. The data we used are the number of DHF patients in a hospital in West Java Indonesia. Most of previous researches used continuous time series model to predict the number of DHF cases because their data range are quite wide. While the data used in this paper have smaller range, then the suitable time series model is discrete time series, specifically integer-valued autoregressive (INAR) order 1 model with Poisson marginal distribution. As a comparison, we also model the number of DHF patients using Markov prediction model (MPM). The MPM model is the probabilistic model which use the concept of the probability of an event occurred based on the previous event. The two models are compared using three kinds of measures; MSE, MAE, and MAPE.

The paper is organized as follows. In Section 2, we provide the theoretical basis of INAR(1)-Poisson. MPM model will be provided in Section 3. In Section 4, the measures of comparison for each model will be given. Then the data are modelled and analysed in Section 5 which divided into the modelling and comparing steps. In Section 6, we will draw a conclusion based on the analysis result.

2. INAR(1)-Poisson model

INAR(1) model is defined as follows [9]:

\[ X_t = \alpha \circ X_{t-1} + \epsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots \] (1)

where \( \alpha \in [0,1] \) and \( \epsilon_t \) is a sequence of uncorrelated non-negative integer-valued random variables having mean \( \mu \) and finite variance \( \sigma^2 \) (follows any discrete distribution).

The INAR(1) model defined in Eq.(1) states that the total individuals at time \( t \), \( X_t \), are the summation of the survivors of the individuals at time \( t - 1 \), \( X_{t-1} \), each with probability of survival \( \alpha \) and the individuals entered the system in interval \((t - 1, t]\), \( \epsilon_t \).

Operator \( \circ \) in Eq.(1) is defined as [10]:

\[ \alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} W_i \] (2)

where \( W_i \sim Bernoulli(\alpha) \).

In this paper, the marginal distribution of \( \epsilon_t \) is Poisson distribution, and the model so called INAR(1)-Poisson. Al-Osh and Alzaid (1987) state that \( X_t \) has a Poisson distribution if and only if \( \epsilon_t \) has a Poisson distribution. If \( \epsilon_t \sim Pois(\lambda(1 - \alpha)) \), using probability generating function technique of Eqn.(1) and discrete self-decomposable properties [10], then \( X_t \sim Pois(\lambda) \).

There are some techniques to estimates the parameters of INAR(1)-Poisson model, \( \alpha \) and \( \lambda \). In this paper, we estimate the parameters using Yule-Walker and conditional least square (CLS) estimation.

2.1 Yule-Walker estimators

The Yule-Walker estimators can be obtained by replacing the covariance at lag \( k \), \( \gamma(k) \), with the sample autocovariance function defined by [9]

\[ \gamma(k) = \alpha^k \text{var}(X_{t-k}) + \sum_{j=0}^{k-1} \alpha^j \text{cov}(X_{t-k}, \epsilon_{t-j}) = \alpha^k \gamma(0) \] (3)

and solve for \( \alpha \) to obtain

\[ \hat{\alpha} = \frac{\sum_{t=0}^{n-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=0}^{n}(x_t - \bar{x})^2} \] (4)

where \( \bar{x} \) is the sample mean. To estimate for \( \lambda \), calculate \( \hat{\epsilon}_t = x_t - \hat{\alpha} x_{t-1} \) for \( t = 1, 2, \ldots, n \) where \( X_t \sim Pois(\lambda) \) and obtained
\[ \hat{\lambda} = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t / \sum_{t=1}^{n} \epsilon_t \]  

(5)

2.2 CLS estimators

The conditional mean of \( X_t \) given \( X_{t-1} \) is given by [9]

\[ E(X_t | X_{t-1}) = \alpha X_{t-1} + \lambda = g(\theta, X_{t-1}) \]  

(6)

where \( \theta = (\alpha, \lambda) \) is the set of parameters to be estimated.

The CLS estimation is derived from the minimization of the sum of squared deviations of the conditional expectation in Eq.(6), that is find the value of \( \alpha \) and \( \lambda \) which minimize

\[ Q_n(\theta) = \sum_{t=1}^{n} [(X_t - g(\theta, X_{t-1}))^2] \]  

with respect to \( \theta \). From the derivation, the estimators of \( \alpha \) and \( \lambda \) are obtained

\[ \hat{\alpha} = \frac{\sum_{t=1}^{n} X_t X_{t-1} - \sum_{t=1}^{n} X_t \sum_{t=1}^{n} X_{t-1}}{n} \]  

(8)

and

\[ \hat{\lambda} = \frac{1}{n} (\sum_{t=1}^{n} X_t - \hat{\alpha} \sum_{t=1}^{n} X_{t-1}) \]  

(9)

3. Markov prediction model

Analogue with INAR(1)-Poisson model, Markov prediction model (MPM) can also be used to model the data at time \( t \) which depend on the data at time \( t - 1 \). Suppose \( \{X(t), t = 0, 1, 2, \ldots \} \) is a stochastic process that takes on a finite number of possible value. If \( X(t) = i \) means that the process is in state \( i \) at time \( t \), then we can calculate the probability of \( X(t + 1) = j | X(t) = i \), that is the probability that the process will be in state \( j \) at time \( t + 1 \) given the previous process is in state \( i \) at time \( t \). Suppose that

\[ P(X(t + 1) = j | X(t) = i, X(t - 1) = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0) = P_{ij} \]  

(10)

for all state \( i_0, i_1, \ldots, i_{t-1}, i, j \) and all \( n \geq 0 \). These stochastic process is known as a Markov chain [11].

\( P_{ij} \) in Eq.(10) has the properties as follows

\[ P_{ij} \geq 0, \quad i, j \geq 0; \quad \sum_{j=0}^{\infty} P_{ij} = 1, \quad i = 0, 1, \ldots \]

Let \( P \) denote the matrix of one-step transition probabilities \( P_{ij} \), so that

\[ P = \begin{pmatrix} P_{00} & P_{01} & \ldots \\ P_{10} & P_{11} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \]  

(11)

Then define the \( n \)-step transition probabilities \( P^n_{ij} \) to be the probability that the process in state \( i \) will be in state \( j \) after \( n \) additional transition. To compute the \( n \)-step transition probability, we can use Chapman-Kolmogorov equations,
\[ P_{ij}^{n+m} = \sum_{k=0}^{\infty} p_{ik}^n p_{kj}^m \text{, for all } n, m \geq 0, \text{all } i, j \] (12)

Suppose \( X(t) = [0 \ 1 \ 2 \ 3 \ \ldots \ \ c] \) is the set of the number of DHF patients, where \( c \) is the maximum number of DHF patients, analogues with the Markov prediction model which was developed by Ren et al. (2015) [12], the MPM for this case is as follows

\[
X(1) = X(0) \cdot P \\
X(2) = X(1) \cdot P \\
\vdots \\
X(t + 1) = X(t) \cdot P
\] (13)

where \( P \) is the transition probability matrix defined by Eq.(11). Using Eq.(13), we can predict the number of DHF patients at time \( t + 1 \) given by the number of DHF patients at time \( t \).

4. MSE, MAE, and MAPE

There are many kinds of measure which can be used to determine the best model of a certain data. In this paper, we will use mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to compare and choose the best model. The formula of three measures are given by

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (\hat{X}_t - X_t)^2
\] (14)

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |\hat{X}_t - X_t|
\] (15)

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{X_t - \hat{X}_t}{X_t} \right| \cdot 100
\] (16)

5. Result and analysis

Data used in this paper is the number of DHF patients in a certain hospital in West Java Indonesia daily from December 27th 2014 to August 24th 2015. Suppose that \( X(t) \) defines the number of DHF patients at time \( t \). The description of the data is shown in figure 1

![The number of DHF patients](image)

**Figure 1.** The number of DHF patients
From figure 1, we can see that the number of DHF patients in those hospital are in range [1,10]. Then we can define that $c = 10$. The maximum number of DHF patients is reached in March 24th 2015.

First step to modeling a data using INAR(1)-Poisson is checking the distribution and PACF of the data. In testing the hypothesis of the data distribution, we use level of significance $\alpha = 5\%$. $H_0$ state that the data follows Poisson distribution and $H_1$ state that data did not follow Poisson distribution. $H_0$ will be rejected if the value of $\text{Sig.} < \alpha$. Based on KS test, we obtain the value of $\text{Sig.} = 0.076 > \alpha$, it means that $H_0$ did not rejected so that the data follows Poisson distribution. Figure 2 shows the partial autocorrelation function (PACF) to determine the order of the INAR model.

![Figure 2. PACF of DHF patients](image)

Figure 2 shows that the lag of the data of DHF patients is in lag 1, so the INAR model for this data is INAR order 1. Table 1 shows the estimation parameter of INAR(1)-Poisson model using CLS and Yule-Walker estimation.

| Parameters       | CLS estimation | Yule-Walker estimation |
|------------------|----------------|------------------------|
| $\hat{\lambda}$ | 0.6530         | 0.6542                 |
| $\hat{\alpha}$  | 0.7904         | 0.7892                 |

Based on the estimation results, we have two INAR(1)-Poisson models that are

$$X_t = 0.7904 \cdot X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{Poi}(0.6530)$$

$$X_t = 0.7892 \cdot X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{Poi}(0.6542)$$

In predicting the number of DHF patients at time $t$ given the number of patients at time $t - 1$, we have to build the transition probability matrix $P$ of the data. Suppose that $X(t)$ define the number of DHF patients at time $t$. Based on figure 1, it is known that the number of DHF patients is in range [1,10] so we have 10 states. $P(X(t) = j|X(t - 1) = i)$ is the probability that the number of DHF patients is $j$ at time $t$ given by the number of patients is $i$ at time $t - 1$. The transition probability matrix $P$ obtained from the calculation is given by

$$P = \begin{bmatrix}
0.5143 & 0.3714 & 0.1143 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.1605 & 0.6173 & 0.1728 & 0.0247 & 0.0247 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$
The prediction of the number of DHF patients then being predicted using Eq.(13) where \( X(t) = [1 \ 2 \ 3 \ \ldots \ 10] \). Figure 3 shows the prediction of the number of DHF patients using INAR(1)-Poisson model with CLS and Yule-Walker estimation and using MPM.

![Graph showing prediction data using INAR(1)-Poisson with CLS and Yule-Walker estimation and using MPM compared to actual value.]

**Table 2.** MSE, MAE, and MAPE of each model

| Model                  | MSE       | MAE        | MAPE     |
|------------------------|-----------|------------|----------|
| INAR(1)-Poisson CLS    | 4.784232365 | 1.70594191 | 70.03029704 % |
| INAR(1)-Poisson YW     | 4.460580913 | 1.58921168 | 66.00342488 % |
| MPM                    | 5.3760102  | 1.57491237 | 48.8855449 %   |

Figure 3. Prediction data using INAR(1)-Poisson with CLS and Yule-Walker estimation and using MPM compared to actual value.
Based on the value of MSE, the best model to predict the number of DHF patients is INAR(1)-Poisson with Yule-Walker estimation. While using MAE and MAPE, the best model is MPM. In general, we can conclude that MPM has better result in predicting the number of DHF patients. The result is rather different if we compare it with the graphic in figure 3. It may happen because the prediction data of MPM which have further distance to the actual value may only a few points. While the other point are closer to the actual value. Otherwise, the INAR(1)-Poisson model may have more points which did not close with the actual data although the distance is not so far away.

6. Conclusion
We have compared the prediction model of the number of DHF patients in a hospital in West Java Indonesia using three models; INAR(1)-Poisson with CLS estimation, INAR(1)-Poisson with Yule-Walker estimation, and Markov prediction model (MPM). Based on the result, MPM has the smallest value of MAE and MAPE compared to the other models although it has the biggest value of MSE. We conclude that, in general, MPM is the best model to predict the number of DHF patients.

7. References
[1] Kumar VS, Mani P, Banstin TMM, Ravikumar G. 2012. Mosquito Larvicidal, Oviposition deterrent and Repellent properties of Acalypha indica L extracts against Aedes aegypti, Anopheles stephensi, and Culex quinquefasciatus. Int J Med Biosci 1(3):33–41.
[2] Lidia K, Setianingrum ELS. 2008. Deteksi dini resistensi nyamuk Aedes albopictus terhadap insektisida organofosfat di daerah endemis demam berdarah di Palu (Sulawesi tengah). MKM. 3:2.
[3] Basker P., Kannan P., Porkaiapandian R.T., Saravanan S. Sridharan S., and Khadiresan M. 2013. Study on entomological surveillance and its significance during a dengue outbreak in the district of Tirunelvi in Tamil Nandu, India. Osong Public Health Res Perspect. 2013.4(3):152-158.
[4] WHO. 1999. Demam berdarah dengue: diagnosis, pengobatan, pencegahan dan pengendalian, edisi 2. WHO.Genewa.
[5] Promprou, S., Jaroensutasinee, M., and Jaroensutasinee, K. 2006. Forecasting dengue haemorrhagic fever cases in southern Thailand using ARIMA models. Dengue Bulletin, Vol. 30, 2006.
[6] Choudhury, M.A.H.Z., Banu, S., and Islam, M.A. 2008. Forecasting dengue incidence in Dhaka, Bangladesh: A time series analysis. Dengue Bulletin, Vol. 32, 2008.
[7] Martinez, E.Z., da Silva, E.A.S., and Fabbro, A.L.D. 2011. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of Sao Paulo, Brazil. Revista da Sociedade Brasileira de Medicina Tropical 44(4); 436-440, Jul-Ago, 2011.
[8] Sriwattanapongse, W. and Khanabsakdi, S. 2011. Modeling and forecasting malaria and dengue hemorrhagic fever incidence and prevalence in Northern Thailand. Journal of Mathematics and System Science I (2011) 52-59.
[9] Al-Osh, M.A., and Alzaid, A.A. 1987. First order integer valued autoregressive (INAR(1)) process. Journal of Time Series Analysis 8: 261-275.
[10] Steutel, F.W., and Harn, K.V. 1979. Discrete analogues of self-decomposability and stability. The Annals of Probability, Vol. 7, No. 5, 893-899.
[11] Ross, S.M. 2007. Introduction to Probability Models, 9th Edition. USA: Academic Press.
[12] Ren, X., Liu, Q., and Zhang, Y. 2015. The proportion of energy consumption structure prediction based on markov chain. Journal of Chemical and Pharmaceutical Research, 2015, 7(3): 378-382.

Acknowledgments
A sincere gratitude goes to the Academic Development Board of Universitas Islam Indonesia, Yogyakarta Indonesia, who has given financial assistance for publication of this research.