A time series model: First-order integer-valued autoregressive (INAR(1))

Cite as: AIP Conference Proceedings **1862**, 030157 (2017); https://doi.org/10.1063/1.4991261
Published Online: 10 July 2017

D. M. Simarmata, F. Novkaniza and Y. Widyaningsih

A Time Series Model: First-order Integer-valued Autoregressive (INAR(1))

D. M. Simarmata, F. Novkaniza[a)], and Y. Widyaningsih

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

[a)]Corresponding author: fevi.novkaniza@sci.ui.ac.id

Abstract. Nonnegative integer-valued time series arises in many applications. A time series model: first-order Integer-valued AutoRegressive (INAR(1)) is constructed by binomial thinning operator to model nonnegative integer-valued time series. INAR (1) depends on one period from the process before. The parameter of the model can be estimated by Conditional Least Squares (CLS). Specification of INAR(1) is following the specification of (AR(1)). Forecasting in INAR(1) uses median or Bayesian forecasting methodology. Median forecasting methodology obtains integer s, which is cumulative density function (CDF) until s, is more than or equal to 0.5. Bayesian forecasting methodology forecasts h-step-ahead of generating the parameter of the model and parameter of innovation term using Adaptive Rejection Metropolis Sampling within Gibbs sampling (ARMS), then finding the least integer s, where CDF until s is more than or equal to u. u is a value taken from the $Uniform(0,1)$ distribution. INAR(1) is applied on pneumonia case in Penjaringan, Jakarta Utara, January 2008 until April 2016 monthly.

INTRODUCTION

Count data is data which record how much the interesting event has occured. Count data recorded by a nonnegative integer (0, 1, 2, ...). Time series of count data arises in many application. If we model such time series using ARIMA (AutoRegressive Moving Average), we will get a continuous number for forecast value. For the example, time series of pneumonia case in Penjaringan, Jakarta Utara, which recorded on January 2008 until April 2016. If we model AR(1) to pneumonia case, we will get 8.94952 number of people who infected in May 2016. It is impossible to represented people by continuous number. Therefore, we need a time series model of count data which give count data for forecast h-step-ahead value.

FIRST-ORDER INTEGER-VALUED AUTOREGRESSIVE (INAR(1)) PROCESS

The operator that used in INAR(1) model is not the same as the multiplying operator. Definition 1 [1] discusses the operator of the model.

Definition 1. Let Z is a nonnegative integer valued variable random, then for any $\alpha \in [0,1]$, binomial thinning operator, which denoted by '$\circ$', is defined as Equation 1:

$$\alpha \circ Z = \sum_{i=1}^{Z} B_i \quad (1)$$
,

where B_i is a series of variable random iid, B_i independent with Z, and

$$\Pr(B_i = 1) = 1 - \Pr(B_i = 0) = \alpha$$

International Symposium on Current Progress in Mathematics and Sciences 2016 (ISCPMS 2016)
AIP Conf. Proc. 1862, 030157-1–030157-8; doi: 10.1063/1.4991261
Published by AIP Publishing. 978-0-7354-1536-2/$30.00

Definition 1 implies that binomial thinning operator generates Z times of Bernoulli trial if the value Z was given. Furthermore, $\alpha \circ Z \mid Z \sim Binomial(Z,\alpha)$. The properties of binomial thinning operator have been written by Silva [2].

Definition of INAR(1) model is written on Definition 2, where assumptions belows must be satisfied:

- $\alpha \in (0,1)$.
- $\alpha \circ Z_{t-1} = \sum_{i=1}^{Z_{t-1}} B_i$, with $B_i \sim Bernoulli(\alpha)$.
- $\{\varepsilon_t\}$ is a sequence of nonnegative integer variable random iid, with mean: $E[\varepsilon_t] = \mu_\varepsilon$ and variance: $Var(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$.

Definition 2. The process $\{Z_t : t = 0, \pm 1, \pm 2, ...\}$ defined as INAR(1) if statisfies Equation 2:

$$Z_t = \alpha \circ Z_{t-1} + \varepsilon_t, \tag{2}$$

The interpretation of INAR(1) model is that the process at time t, that is Z_t, is the summation of the survivors at $t-1$ that can survive until t with probability of surviving α and the objects which entered the system in the time interval $(t-1, t]$ which denoted by ε_t [3]

Characteristics of Model INAR(1)

This subsection discusses mean, variance, autocovariance, autocorrelation, and partial autocorrelation function for INAR(1) model. The autocorrelation and partial autocorrelation function can be considered as specification tools of the model.

Expected function for model INAR(1) can be written as:

- Conditional expectation of Z_t, given $Z_{t-1}: E(Z_t \mid Z_{t-1}) = \alpha Z_{t-1} + \mu_\varepsilon$.
- Unconditional expectation of Z_t: $E(Z_t) = \alpha^t E(Z_0) + \mu_\varepsilon \sum_{j=0}^{t-1} \alpha^j$.
- Variance of Z_t: $Var(Z_t) = \alpha^{2t} Var(Z_0) + (1-\alpha) \sum_{j=1}^{t} \alpha^{2j-1} E(Z_{t-j}) + \sigma_\varepsilon^2 \sum_{j=1}^{t} \alpha^{2(j-1)}$.
- Autocovariance of Z_t: $Cov(Z_t, Z_{t-k}) = \alpha^k Var(Z_t)$.
- Autocorrelation of Z_t: $Corr(Z_t, Z_{t-k}) = \alpha^k$.

The magnitude of autocorrelation function (ACF) decreases exponentially as the number of lags k increases. We get the partial autocorrelation function (PACF) if the value of ACF is substituted to the following Equation 3 and 4:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\, \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\, \rho_j}, \tag{3}$$

where

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{k,k}\, \phi_{k-1,k-j}, \tag{4}$$

for $j = 1, 2, ..., k-1$. Because ACF of INAR(1) looks like ACF of AR(1), then we obtain that PACF of INAR(1) also looks like ACF AR(1). Therefore, the only specification of INAR(1) has PACF significance on lag 1.

Parameter Estimation using Conditional Least Squares (CLS) Method

The CLS method is finding parameter estimation by minimizing sum square of difference between Z_t and conditional expectation of Z_t, given Z_{t-1} [4]. With the assumption $\varepsilon_t \sim Poisson(\lambda)$ and given time series until $t = n$ we obtain that:

$$\hat{\alpha} = \frac{\sum_{t=1}^{n} Z_t Z_{t-1} - \hat{\lambda} \sum_{t=1}^{n} Z_{t-1}}{\sum_{t=1}^{n} Z_{t-1}^2} \tag{5}$$

and

$$\hat{\lambda} = \frac{1}{n}\left(\sum_{t=1}^{n} Z_t - \hat{\alpha} \sum_{t=1}^{n} Z_{t-1}\right). \tag{6}$$

Substitution Equation 6 to Equation 5, to obtain parameter estimation based on time series, that is (Equation 7 and 8):

$$\hat{\alpha} = \frac{\sum_{t=1}^{n} Z_t Z_{t-1} - \frac{1}{n} \sum_{t=1}^{n} Z_t \sum_{t=1}^{n} Z_{t-1}}{\sum_{t=1}^{n} Z_{t-1}^2 - \frac{1}{n}\left(\sum_{t=1}^{n} Z_{t-1}\right)^2} \tag{7}$$

and

$$\hat{\lambda} = \frac{1}{n}\left(\frac{\sum_{t=1}^{n} Z_t \sum_{t=1}^{n} Z_{t-1}^2 - \sum_{t=1}^{n} Z_t Z_{t-1} \sum_{t=1}^{n} Z_{t-1}}{\sum_{t=1}^{n} Z_{t-1}^2 - \frac{1}{n}\left(\sum_{t=1}^{n} Z_{t-1}\right)^2}\right). \tag{8}$$

Diagnostic Model

After obtaining parameter estimation of model INAR(1), we need to diagnose the model. Residual of fitted model is not has correlation each other. Residual for model INAR(1) is $r_t = Z_t - \alpha Z_{t-1} - \lambda$. If the model is adequate, then the plot of standardized residual scatter around a zero horizontal level with no trends [4].

Forecasting Method

If we use the conditional expectation concept to forecast INAR(1) model, will give us continuous number forecast value. Therefore, we use other method in order to get the nonnegative integer forecast value. There are two methods that will be displayed here, that are, Median forecasting method [5] and Bayesian forecasting method [6].

Median Forecasting Method

Median forecasting method as the finding of value which expected absolute error minimum. Expected absolute error is the difference between expected value Z_t and the real value Z_t, given the present value. Median forecast method tells that forecast value that minimize expected absolute error minimum is the conditional median given present value. We will define conditional median of Z_{n+h}, given $Z_n = z_n$, as the smallest non-negative integer m_h such that $\sum_{z=0}^{m_h} p(z \mid z_n) \geq 0.5$.

Theorem 1 [5] can be used to construct probability mass function (pmf) for the model INAR(1), given $Z_n = z_n$.

Theorem 1. For the INAR(1) model, with $\varepsilon_t \sim Poisson(\lambda)$, of Z_{t+h} given Z_t is a convolution of $Binomial(Z_n, \alpha^h)$ and $Poisson\left(\lambda \frac{1-\alpha^h}{1-\alpha}\right)$. That is, the h-step-ahead conditional mgf is given by Equation 9:

$$M_{Z_{t+h}|Z_t}(s) = \left[\alpha^h e^s + \left(1-\alpha^h\right)\right]^{Z_t} \exp\left\{\lambda \frac{1-\alpha^h}{1-\alpha}\left(e^s - 1\right)\right\} \tag{9}$$

From Theorem 1, we obtain pmf of INAR(1) model for value Z_{t+h}, given Z_t that is (Equation 10):

$$p_h(z \mid Z_t) = \sum_{i=0}^{\min(z,Z_t)} \binom{Z_t}{i} \left(\alpha^h\right)^i \left(1-\alpha^h\right)^{Z_t - i} \frac{\exp\left\{-\lambda \frac{1-\alpha^h}{1-\alpha}\right\}}{(z-i)!} \left(\lambda \frac{1-\alpha^h}{1-\alpha}\right)^{z-i}. \tag{10}$$

This method is not suitable for the condition which $p_h(0 \mid z_n) > 0.5$ because the median will not be defined.

Bayesian Forecasting Method

The idea of Bayesian forecasting method [6] is based on the two terms of INAR(1) model are variable random. Hence, the conjugate prior of α is $Beta(a,b)$, while the conjugate prior of λ is $Gamma(c,d)$.

The full conditional posterior distribution of α written as Equation 11:

$$\pi(\alpha \mid \lambda, Z) \propto \alpha^{a-1}(1-\alpha)^{b-1} \prod_{t=2}^{n} \sum_{i=0}^{M_t} \frac{\lambda^{Z_t - i}}{(Z_t - i)!} \binom{Z_{t-1}}{i} \alpha^i (1-\alpha)^{Z_t - i} \tag{11}$$

,

while the full conditional posterior distribution of λ written as Equation 12:

$$\pi(\lambda \mid \alpha, Z) \propto \lambda^{c-1} \exp(-(d+n-1)\lambda) \prod_{t=2}^{n} \sum_{i=0}^{M_t} \frac{\lambda^{Z_t - i}}{(Z_t - i)!} \binom{Z_{t-1}}{i} \alpha^i (1-\alpha)^{Z_t - i}. \tag{12}$$

The predictive posterior Z_{n+h}, given Z_n, is complicated. Hence, we can not find the solution using the standard method of Bayesian. The algorithm of Bayesian forecasting method can be used to find the forecasting value.

The Algorithm of Bayesian forecasting method [6] written as follows:

1. Finding estimation parameter of model INAR(1).
2. Defining m as how much the iterations that will be performed and S_n as the sequences of proceeds of a collection (α, λ) in every iteration.
3. Doing Adaptive Rejection Metropolis Sampling (ARMS) in Gibbs sampling procedures, where (Equation 13 and 14):

$$\alpha^{[j]} \sim \pi(\alpha \mid \lambda^{[j-1]}, Z) \tag{13}$$

$$\lambda^{[j]} \sim \pi(\lambda \mid \alpha^{[j-1]}, Z) \tag{14}$$

4. Sampling $u \sim Uniform(0,1)$.
5. Finding nonnegative integer s, such as statisfied Equation 15:

$$\sum_{i=0}^{s} p_h(z \mid Z_t) \geq u \tag{15}$$

6. Obtaining $\hat{Z}_{n+h,i}$.

The Algorithm ARMS in Gibbs Sampling [7, 8] written as follows:

1. Initializing m as number of iteration, S_n as the collection that contain sequences of α that has been generated, and α_{CLS} as the estimation of α that has been obtained from CLS method.
2. Sampling α^* from the probability posterior sampling.
3. Sampling $u \sim Uniform(0,1)$.
4. If $u > \pi(\alpha^* \mid \lambda, Z) / \exp h_n(\alpha^* \mid \lambda, Z)$, then α^* enter S_n, $n \leftarrow n+1$, and back to step 2. If not, then go to next step.
5. Sampling $u \sim Uniform(0,1)$.

6. If $u > \min\left[1, \dfrac{\pi(\alpha^* \mid \lambda, Z)\min\{\pi(\alpha_0 \mid \lambda, Z), \exp h_n(\alpha_0 \mid \lambda, Z)\}}{\pi(\alpha_0 \mid \lambda, Z)\min\{\pi(\alpha^* \mid \lambda, Z), \exp h_n(\alpha^* \mid \lambda, Z)\}}\right]$, then α_0 enter S_n, where α_0 is taken from the step before and has entered S_n, then $n \leftarrow n+1$. If not, α^* enter S_n and $n \leftarrow n+1$.

APPLICATION ON PNEUMONIA CASES IN PENJARINGAN

In this section, we will use all theories and concepts of INAR(1) that has been discussed before. We use Penjaringan's affected pneumonia population data on January 2008 until April 2016 monthly. All the plot, graphics, and calculations use statistical program: R x64 3.3.1.

According to CDC [8], pneumonia is an infection of the lungs that is still the leading cause of death in children younger than 5 years old worldwide. This is can be suffered by people in all ages. Pneumonia can be caused by viruses, bacteria, and fungi. To prevent pneumonia and other respiratory infection is by washing hands regularly, taking good care of medical problems, and quitting smoking. Besides that, pneumonia can be prevented with vaccines and treated with medicine, depending on the cause.

Penjaringan, located in Jakarta Utara, is one of the historical region in Jakarta. We work with Penjaringan's population of people who affected pneumonia on January 2008 until April 2016. Using Augmented Dickey-Fuller (ADF) test, the time series is stationary.

ACF and PACF plots on Fig. 1 display us that the current value of the series Z_t is a linear combination of the one most recent past values of itself plus an innovation terms that contains everything new in the series at time t that is not explained by the past values. If we consider the series as AR(1), then it will be modeled by:

$$Z_t = 0.4257\, Z_{t-1} + a_t,$$

where $a_t \sim Normal(0, \hat{\sigma}_a^2 = 13.71)$ and $\hat{\mu} = 5.9471$.

The forecast value of h step period ahead for AR(1) are:

$\hat{Z}_{100}(1) = 8.94952$, $\hat{Z}_{100}(2) = 7.22523$, $\hat{Z}_{100}(3) = 6.4912$, and $\hat{Z}_{100}(4) = 6.178723$.

At a glance, the forecast with AR(1) model tells that 8.94952 people will affect pneumonia on May 2016. These results cannot be represented the real condition because the forecast value is continuous number. Inspite of the number of people is only represented by count data, we need to model Penjaringan's affected pneumonia population using INAR(1).

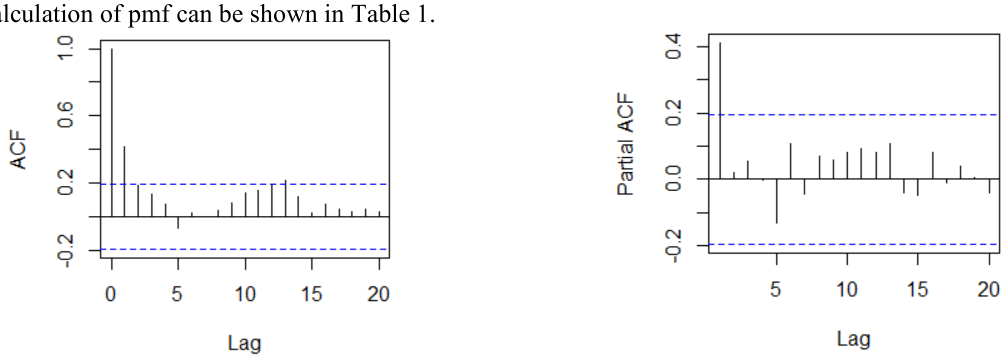
Parameter estimation with CLS gives the INAR(1) model for Pneumonia case in Penjaringan as:

$$Z_t = 0.4378081 \circ Z_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim Poisson(\hat{\lambda} = 3.339469)$. Then, we have to plot the standardized residual of model. Figure 2 is the standardized residual plot of Penjaringan's affected pneumonia population.

Median Forecast Method

The calculation of pmf can be shown in Table 1.



(a) (b)

FIGURE 1. (a) ACF and (b) PACF of time series of Penjaringan's population of people who affected pneumonia on January 2008 until April 2016

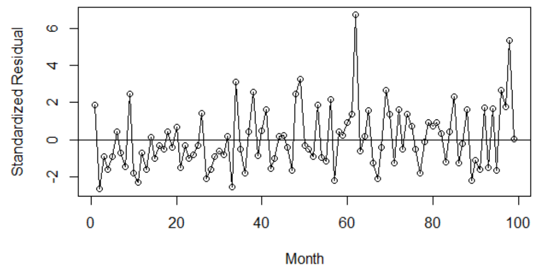


FIGURE 2. Standardized residual plot of time series of Penjaringan's affected pneumonia population

TABLE 1. Probability mass function calculation result for time series of Penjaringan's population who affected pneumonia (Januari 2008 – April 2016)

h	1	2	3	4
$p_h(0 \mid z_n)$	0.000019869	0.00051682	0.0013864	0.0020123
$p_h(1 \mid z_n)$	0.00026751	0.0040747	0.0091953	0.012512
$p_h(2 \mid z_n)$	0.0017224	0.015874	0.030418	0.038879
$p_h(3 \mid z_n)$	0.0070675	0.040755	0.066919	0.080501
$p_h(4 \mid z_n)$	0.02078	0.0776	0.11015	0.12495
$p_h(5 \mid z_n)$	0.046683	0.11692	0.14472	0.15508
$p_h(6 \mid z_n)$	0.083462	0.14526	0.15808	0.16032
$p_h(7 \mid z_n)$	0.12216	0.1531	0.14767	0.142
$p_h(8 \mid z_n)$	0.14947	0.13979	0.12044	0.11
$p_h(9 \mid z_n)$	0.15543	0.11236	0.087129	0.075706
$p_h(10 \mid z_n)$	0.13919	0.080525	0.056608	0.046874
$p_h(11 \mid z_n)$	0.10857	0.051987	0.033366	0.026372
$p_h(12 \mid z_n)$	0.074469	0.030496	0.017991	0.013595
$p_h(13 \mid z_n)$	0.045309	0.016373	0.0089371	0.0064662
$p_h(14 \mid z_n)$	0.024641	0.0080956	0.0041143	0.0028547
$p_h(15 \mid z_n)$	0.012062	0.0037062	0.0017644	0.0011758
Median	9 (0.594082279)	7 (0.55410052)	6 (0.5208687)	6 (0.57425)

First, we have to generate the prior distribution of $\alpha \circ Z_t$ and ε_t using ARMS. The convergence diagnostics of Gelman *et al*. [9] shows that 5000 iterations indicate convergence of both posterior distributions. The trace plots show in Fig. 3.

Then, we obtain h -step-ahead forecast value using Bayesian forecasting method in the Table 2.

Using accuracy test, Median minimizes Mean Squares Estimation (MSE), Mean Absolute Estimation (MAE), and Mean Absolute Percentage Estimation (MAPE). On h =1 the forecast value for two methodology is different. So, if we compare accuracy test of two forecasting methodology result, we will found that the result of Median forecasting methodology minimizes MSE, MAE, and MAPE. The accuracy test shown in Table 3.

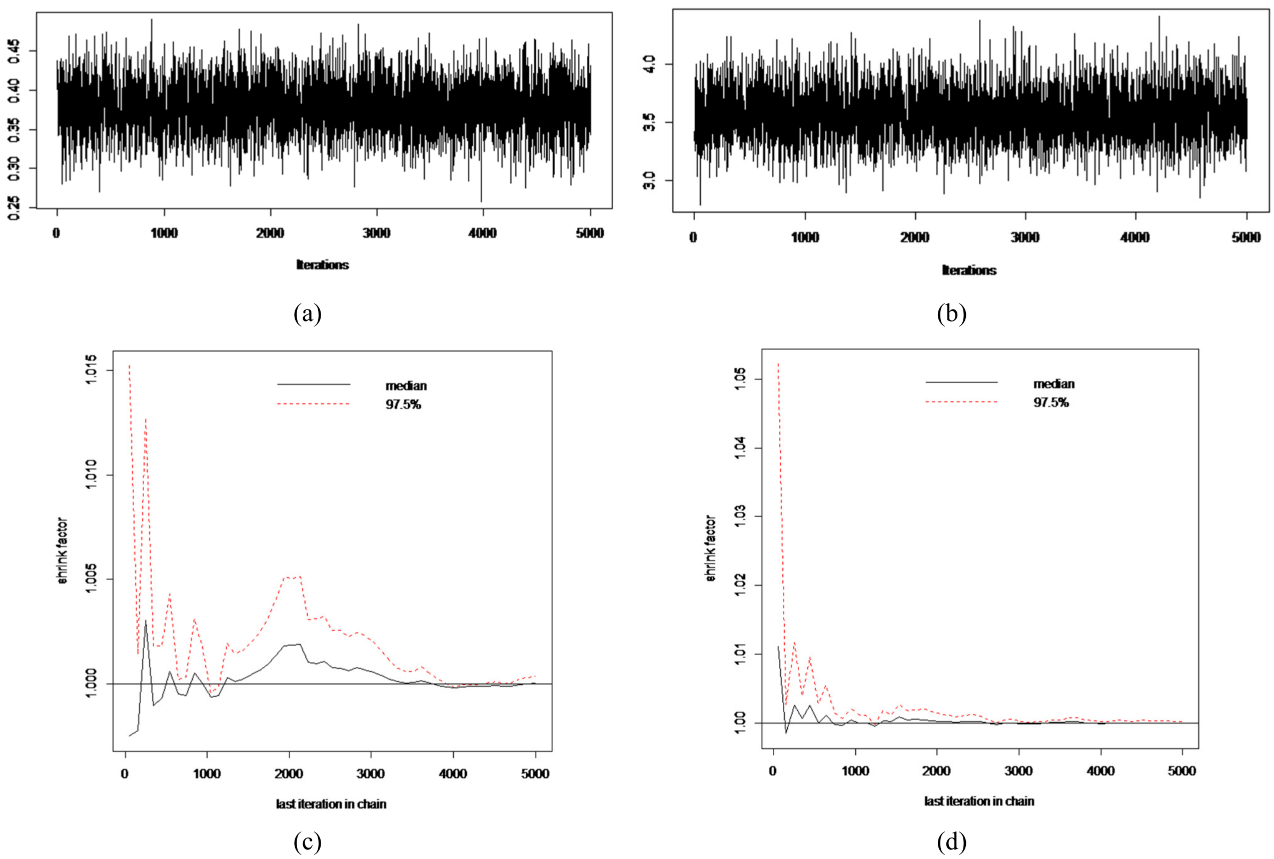


FIGURE 3. Trace plots of (a) α , (b) λ , and plot of shrink factor of generating iterations of (c) α , (d) λ

TABLE 2. h -step-ahead forecast result using Bayesian forecasting method

h	**1**	**2**	**3**	**4**
Mean	8.5182	6.7966	6.1598	5.876
Median	8	7	6	6
Mode	9	6	6	5

TABLE 3. MSE, MAE, MAPE calculation results

Method	**MSE**	**MAE**	**MAPE**
Bayes (Mode)	69.75	7.25	1.211012
Bayes (Median)	61.25	6.75	1.01587
Median	63.5	7	1.061012

ACKNOWLEDGMENTS

Authors give thankful for referees, because of their generosity to share their knowledge and Dinas Kesehatan DKI Jakarta to share data that shown on Application on Pneumonia case in Penjaringan. Authors receive comment from the reader by email.

REFERENCES

1. F. W. Steutel and K. V. Harn, Ann. Probab. **7**, 893–899 (1979).
2. I. M. M. Silva, "Contributions to the Analysis of Discrete-Valued Time Series," M.Sc. thesis, Universidade do Porto, Porto, 2005.
3. M. A. Al-Osh and A. A. Alzaid, J. Time Ser. Anal. **8**, 261–275 (1987).
4. J. D. Cryer and K. S. Chan, *Time Series Analysis: with Application in R* (Springer, Berlin, 2010).
5. R. K. Freeland, Statistical Analysis of Discrete Time Series with Application to the Analysis of Workers Compensation Claims Data, Ph.D thesis, The University of British Columbia, Vancouver, 1998.
6. N. Silva, I. Pereira, and M. E. Silva, Stat. J. **7**, 119–134 (2009).
7. W. R. Gilks, N. G. Best, and K. K. C. Tan, Appl. Stat. **44**, 455–472 (1994).
8. W. R. Gilks and P. Wild, Appl. Stat. **41**, 337–348 (1992).
9. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis : Text in Statistical Science* (Chapman & Hall/CRC, New York, 2000).