Forecasting the Number of Prisoners in Nganjuk With Integer-Valued Pth-Order Autoregressive (INAR(P))

Bella Belinda* and Mila Novita
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

*E-mail: bella.belinda@sci.ui.ac.id

Abstract. The most used time series model is the time series model which assumes the variables being tested are continuous in a discrete time period and produces continuous values. Whereas in many applications, such as number of monthly accidents, number of doctor visits per year a person makes, etc., needed a discrete time series model to handle discrete variables and produce discrete values as well. Time series model that handles count or non-negative integer data is the Integer-valued Autoregressive model with the pth-order or INAR(p). INAR(p) is the counterpart of AR(p) model for integer data. To get integer results, this model uses binomial thinning operator which implements probabilistic operations with discrete distribution that are suitable to model count data such as Poisson and Binomial, also use median forecasting method. Model parameters will be estimated using the Yule-Walker method. In this research, INAR(p) time series model will be applied to number of prisoners in Nganjuk from April 2013 until July 2016 to help tackle overcapacity problems in prisons that led to many negative impacts. Through model specification, the best model for forecast the case is INAR(2). In this data, based on the measure of AIC, BIC, and AICc, the INAR(2) model achieved better performance than its AR(2) counterpart.

1. Introduction
Time series model can be either continuous or discrete based on time period. In a time series model with continuous time period, the sequence of observations is recorded in continuous time. Whereas in a time series with discrete time period, the sequence of observations is recorded at a specific time and the same distance, for example per hour, daily, weekly, monthly, or yearly with the measured variable assumed in the form of continuous or discrete random variable. A popular and widely used time series model is time series model with discrete time period and assumes the variables being tested are continuous, that is ARIMA (Autoregressive Integrated Moving Average). One of the ARIMA model is AR (Autoregressive) which is a linear model that mostly works on stationary time series.

The discrete time series model that assume the measured variable is a continuous random variable will produce a continuous forecast value. However, in reality many events, objects, or individuals whose observations are recorded at certain consecutive intervals or points of time discretely in non-negative integer \( \{0, 1, 2, \ldots\} \) which is called count data. Some simple examples are the monthly number of prisoners that used in this research, the number of monthly accidents, the number of doctor visits per year a person makes, and so on. Thus, it is not appropriate to use time series models such as ARIMA or AR for these cases because the expected forecast results should be represented in discrete numbers or count data as well.
Therefore, various time series models with discrete random variables were developed. Al-Osh & Alzaid (1990) [1] and Du & Li (1991) [2] developed a time series model for count data using AR(p), Integer-valued Autoregressive (INAR) with pth-order (INAR(p)). This time series model uses the binomial thinning operator introduced by Steutel & van Harn (1979) [3]. Binomial thinning operator denoted 'о'. This operator applies probabilistic operations that overcomes the problem of multiplication in a time series model with a continuous random variable which cannot be applied to the count process because it cannot produce discrete results [4].

In this research, INAR(p) model will be applied to the number of prisoners in Nganjuk from April 2013 until July 2016 in order to get coherent forecast result. This research is conducted to the number of prisoners to tackle overcapacity problem in prisons, especially at Nganjuk, so in the future the number of prisoners can be allocated properly. Then, we will do some model specification to determine the order of the model in Section 2. The parameter of the model will be estimated using Yule-Walker method in Section 3. Then, the data will be forecasted h-step ahead with median forecasting method in Section 5.

2. The data and some preliminaries

The methods exposited and developed in this paper are applied to test data consisting of 40 data of monthly number of prisoners (that have been in prison for 3 months until 1 year) in Nganjuk, East Java which recorded from April 2013 to July 2016 from Central Bureau of Statistics Indonesia.

The problem of overcapacity in prisons in Indonesia has been a scourge for a long time. In Nganjuk itself, in 2018 it was reported that the prisons in the area had almost 100 percent overcapacity. This problem is a serious problem and needs to be addressed immediately because of many negative impacts, for example inconvenience that can lead to emotional problems because the assisted residents or prisoners are crowded, services are not optimal due to the unbalanced ratio of the number of officers and assisted residents, unsustainable sanitation and environmental cleanliness. disease, to the lack of supervision so that it is prone to abuse in prison.

Therefore, it is hoped that with this research predicting the number of prisoners each month can make prisons in Nganjuk better in allocating their assisted residents so that they can overcome this overcapacity problem and provide better services in the future.

The sample mean of the series is 34.2 and the sample variance is 12.1828.

2.1. Stationary Test

The stationary test was carried out by using the Augmented Dickey-Fuller test or ADF test. The ADF statistical value of the data is -3.7958. This value is compared with the $1 - \alpha$ quantile of the ADF statistical table. Since the absolute value of the ADF statistic is greater than the quantile value for $\alpha = 0.05$, 2.88 thus the null hypothesis is not accepted. Then, the time series of the data is a stationary series.
Model specification for INAR($p$) model is following ACF and PACF of AR($p$) model.

![Figure 2. ACF plot of data](image)

![Figure 3. PACF plot of data](image)

From Fig. 3, it shows that for lag more than 2 PACF it is not too significant and tends to be close to 0. In Fig. 2, the ACF plot is slowly decreasing. Thus, PACF is significant up to second order and the time series of the data can be modelled with INAR(2) or AR(2).

3. Method

While common time series like AR($p$) using scalar operation that can’t produce discrete results, INAR($p$) model using binomial thinning operator as the replacement. Binomial thinning operator apply probabilistic operations to $X$ count random variable with a 'thinning' operator '○' that leads to integer results between 0 and $X$. In this section, will be shown how the binomial thinning operator used to build the INAR($p$) and INAR(2) for this case.

3.1. Binomial thinning operator

**Definition 1.** Let $X$ be a discrete random variable with range $\{0, \ldots, n\}$ or $\mathbb{N}$, let $\alpha \in [0, 1)$. Define the random variable

$$\alpha \circ X := \sum_{i=1}^{X} B_i \alpha$$

where the $B_i$ are i.i.d. Bernoulli indicators which also independent of $X$. Then, we could say that $\alpha \circ X$ arises from $X$ Bernoulli trials so that $\alpha \circ X \sim \text{Binomial}(X, \alpha)$ with expectation $E[\alpha \circ X] = \alpha E(X)$ and
variance $V[\alpha o X] = \alpha^2 \sigma_e^2 + \alpha(1 - \alpha)E(X)$. Silva [5] has written other properties of binomial thinning operator.

3.2. Integer-valued $p$th-order autoregressive (INAR($p$)) model

INAR($p$) process $\{X_t; t = 0, \pm 1, \pm 2, \ldots\}$ is defined with:

$$X_t = \alpha_1 o X_{t-1} + \alpha_2 o X_{t-2} + \ldots + \alpha_p o X_{t-p} + \varepsilon_t. \quad (1)$$

It is assumed that the fixed parameter $\alpha \in [0, 1)$ and that the innovations $\varepsilon_t$ are non-negative, independent and identically distributed (i.i.d.) random variables with mean $\mu_\varepsilon$ and variance $\sigma_\varepsilon^2$. It is assumed that the error or innovations is Poisson distributed so, $\mu_\varepsilon = \sigma_\varepsilon^2$. $\alpha_1 o X_{t-1}$, $\alpha_2 o X_{t-2}$, ..., $\alpha_p o X_{t-p}$ assumed to be counted independently of each other.

$X_t$ is the number of individuals, objects, or events at time $t$, $\alpha_p o X_{t-p}$ is the survival term, and $\varepsilon_t$ as error or referred as innovation. Operating binomial thinning produces survival term represents the number of individuals, objects, or events from the Bernoulli experiment that can survive (in other words success) with a survival probability of $\alpha_p$ at time $t - p$ [6].

There are two approaches available for INAR($p$), that is Al-Osh & Alzaid (AA) and Du & Li (DL). In this research, DL approaches is chosen because of most of the literature using Du & Li approach and it is said that Al-Osh & Alzaid approach has more complicated calculations because of its conditional multinomial assumption.

3.3. INAR(2) model for the data

Based on Subsection 2.2, we will use $p = 2$. INAR(2) $\{X_t; t = 0, \pm 1, \pm 2, \ldots\}$ process for the model is defined with:

$$X_t = \alpha_1 o X_{t-1} + \alpha_2 o X_{t-2} + \varepsilon_t. \quad (2)$$

INAR(2) process of Du and Li (1991) denoted as INAR(2)-DL treat the thinning operations $\alpha_1 o X_{t-1}$ and $\alpha_2 o X_{t-2}$ remain independent of the innovation process. However with this approach, the marginal distribution of $X_t$ is no longer Poisson if the innovations are. The INAR(2)-DL is stationary as long as $\alpha_1 + \alpha_2 < 1$ [7]. Also, the correlation properties and other latter's properties of INAR(2)-DL are identical to the linear Gaussian AR(2) model.

4. Parameter Estimation: Yule-Walker Method

The two parameters in INAR(2) model will be estimated with Yule-Walker method [8]. Yule-Walker method is also known as method of moments and widely used to estimating parameters of stationary autoregressive model. This method could be conducted with matrix solution with autocorrelation fuctions, $\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + \ldots + \alpha_p \rho_{k-p}$:

$$\Gamma \alpha = \rho \quad (3)$$

where $\Gamma = [\rho_{|t-j|}]_{p \times p}$; $\alpha = (\alpha_1 \alpha_2 \ldots \alpha_p)_T$; $\rho = (\rho_1 \rho_2 \ldots \rho_p)_T$.

Then, with replacing the theoretical autocorrelation $\rho_k$ in (3) with the sample autocorrelation $r_k$ yields the Yule-Walker parameter estimate for $\alpha$ where:

$$\hat{\Gamma} \hat{\alpha} = \hat{\rho} \quad (4)$$
Based on (4), the Yule-Walker estimators for $\alpha_1$ and $\alpha_2$ are

$$\hat{\alpha}_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2} \quad (5)$$

$$\hat{\alpha}_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \quad (6)$$

We get the estimated value of $\alpha_1$ and $\alpha_2$ are 0.6878 and 0.2319. So, INAR model for forecasting the number of prisoners in Nganjuk is:

$$X_t = 0.6878oX_{t-1} + 0.2319oX_{t-2} + \varepsilon_t. \quad (7)$$

5. Forecasting with INAR models

The method that used to forecast the INAR(2) model is median forecasting by calculating the conditional probability of each possible nonnegative integer value, then selecting a forecast value with a cumulative conditional probability that has value greater than 0.5. This method is used rather than conditional expectation forecast because forecast value that obtained by conditional expectation forecast that usually used in AR model will yield real result rather than integer value [9].

For INAR(2) model, the conditional probability if it is known that $\alpha_1 oX_{t-1}, \alpha_2 oX_{t-2}, \ldots, \alpha_p oX_{t-p}$ has a Binomial distribution and it is assumed that the innovation term has a Poisson distribution [10].

$$P(X_t | X_{t-1}, X_{t-2}) = \sum_{i_1=0}^{\min(X_{t-1}, X_{t-2})} \binom{X_{t-1}}{i_1} \alpha_1^{i_1} (1 - \alpha_1)^{X_{t-1} - i_1} \sum_{i_2=0}^{\min(X_{t-2}, X_{t-1} - i_1)} \binom{X_{t-2}}{i_2} \alpha_2^{i_2} (1 - \alpha_2)^{X_{t-2} - i_2} e^{-\lambda} \frac{\lambda^{X_t - (i_1 + i_2)}}{(X_t - (i_1 + i_2))!} \quad (8)$$

Then, we will conduct $h$-step forecasting with $h = 3$ and get conditional probability and cumulative conditional probability of every possible integer value. From all possible integer value, we only show some values that fall between values that has cumulative conditional probability greater than 0.5.

**Table 1.** Conditional probability of $X_{40+h}$ Given $X_{38+h}$ and $X_{39+h}$ for time series of the number of prisoners in Nganjuk

|                      | $H$       |
|----------------------|-----------|
|                      | 1         | 2         | 3         |
| $p_h(X_{40+h} | X_{38+h}, X_{39+h})$ |           |           |           |
| ...                 | ...       | ...       | ...       |
| $p_h(40|X_{38+h}, X_{39+h})$ | 0.040612  | 0.048278  | 0.056413  |
| $p_h(41|X_{38+h}, X_{39+h})$ | 0.051917  | 0.059991  | 0.067875  |
| $p_h(42|X_{38+h}, X_{39+h})$ | 0.063214  | 0.07094   | 0.077692  |
| $p_h(43|X_{38+h}, X_{39+h})$ | 0.073343  | 0.079861  | 0.084639  |
| $p_h(44|X_{38+h}, X_{39+h})$ | 0.081119  | 0.085624  | 0.087794  |
Based on Table 2, forecast values for 3-step-ahead for the data are 46, 45, and 44. So, in August 2016 there will be 46 assisted resident or prisoners, in September 2016 there will be 45 prisoners, and in October 2016 there will be 44 prisoners. These results also show that INAR model could produce integer results.

5.1. Residual analysis
We also must check whether our residuals or innovations has Poisson distribution. This will be checked with chi-square goodness-of-fit test [11]. The p-value from the test is 0.000856 which concludes that the residuals of data is not Poisson distributed. Ideally, to use the INAR model the residuals need to be Poisson distributed. However, just like the residual assumption that is normally distributed in the AR model, which is not mandatory but important, this data can still be used with the INAR model where the consequence is that the forecast confidence interval will widen.

\[
\begin{align*}
\mathcal{P}_h(45|X_{38+h}, X_{39+h}) & = 0.085562 \quad 0.087466 \quad 0.086744 \\
\mathcal{P}_h(46|X_{38+h}, X_{39+h}) & = 0.086098 \quad 0.08516 \quad 0.081669 \\
\mathcal{P}_h(47|X_{38+h}, X_{39+h}) & = 0.082685 \quad 0.079055 \quad 0.073295 \\
\mathcal{P}_h(48|X_{38+h}, X_{39+h}) & = 0.075811 \quad 0.069996 \quad 0.062728 \\
\mathcal{P}_h(49|X_{38+h}, X_{39+h}) & = 0.066385 \quad 0.059132 \quad 0.051211
\end{align*}
\]

Table 2. Cumulative conditional probability of $X_{40+h}$ Given $X_{38+h}$ and $X_{39+h}$ for time series of the number of prisoners in Nganjuk

| $\mathcal{P}_h(X_{40+h}|X_{38+h}, X_{39+h})$ | 1 | 2 | 3 |
|-----------------------------------------------|---|---|---|
| $\mathcal{P}_h(40|X_{38+h}, X_{39+h})$       | 0.128433 | 0.160209 | 0.19901 |
| $\mathcal{P}_h(41|X_{38+h}, X_{39+h})$       | 0.180349 | 0.220201 | 0.266885 |
| $\mathcal{P}_h(42|X_{38+h}, X_{39+h})$       | 0.243563 | 0.29114 | 0.344577 |
| $\mathcal{P}_h(43|X_{38+h}, X_{39+h})$       | 0.316907 | 0.371001 | 0.429216 |
| $\mathcal{P}_h(44|X_{38+h}, X_{39+h})$       | 0.398026 | 0.456625 | 0.51701 |
| $\mathcal{P}_h(45|X_{38+h}, X_{39+h})$       | 0.483587 | 0.544091 | 0.603754 |
| $\mathcal{P}_h(46|X_{38+h}, X_{39+h})$       | 0.569685 | 0.629251 | 0.685423 |
| $\mathcal{P}_h(47|X_{38+h}, X_{39+h})$       | 0.65237 | 0.708306 | 0.758718 |
| $\mathcal{P}_h(48|X_{38+h}, X_{39+h})$       | 0.728181 | 0.778302 | 0.821446 |
| $\mathcal{P}_h(49|X_{38+h}, X_{39+h})$       | 0.794566 | 0.837434 | 0.872657 |
| ...                                          | ... | ... | ... |
5.2. Model diagnosis
In this section, will be compared between AR(2) and INAR(2) model to see which one is better to model the data using the information criteria, AIC, BIC, and AICc [12]. AICc is used to overcome the bias of AIC when dealing with small samples. Regarding to our data that only consist of 40 data which consider as small sample, then we will focus on the AICc measure.

Table 3. Model diagnosis AR(2) and INAR(2)

| Model   | AIC   | AICc  | BIC   |
|---------|-------|-------|-------|
| AR(2)   | 240.35| 247.1 | 241.49|
| INAR(2) | 161.03| 161.34| 113.87|

Because AICc of INAR(2) is smaller than AICc of AR(2) then we can conclude that INAR(2) is the better model for the number of prisoners in Nganjuk from April 2013 to July 2016 that could handle integer data type.

6. Conclusion
The main contribution and purpose of this paper is to discuss the use of the integer-valued pth-order autoregressive (INAR(p)) model for count time series, in this case INAR(2) model for forecast the number of prisoners (that have been in prison for 3 months until 1 year) in Nganjuk from April 2013 to July 2016. It can be seen that INAR(2) model is the better in model the data which is count data and produce coherent forecast results.

Acknowledgements
This work is supported by Program Publikasi Terindeks Internasional (PUTI) Prosiding 2020 funded by Directorate of Research and Development Universitas Indonesia No. NKB-3610/UN2.RST/HKP.05.00/2020.

References
[1] Alzaid A A and Al-Osh M A 1990 An integer-valued pth-order autoregressive structure (INAR(p)) process Journal of Applied Probability 27 314-323
[2] Du J G and Li Y 1991 The integer-valued autoregressive (INAR(p)) model Journal of Time Series Analysis, 12, 129-142
[3] Steutel F W and van Harn K 1979 Discrete Analogues of Self-Decomposability and Stability The Annals of Probability Vol. 7, No. 5, 893-899
[4] Weiß C H 2018 An Introduction to Discrete-Valued Time Series John Wiley & Sons, Inc Chichester
[5] Silva M E and Oliveira V L 2004 Differential equations for the higher-order moments and cumulants of the INAR(1) Model Journal of Time Series Analysis Vol. 25, No. 3, 317-333
[6] Al-Osh M A and Alzaid A A 1987 First-order integer valued autoregressive (INAR(1)) process Journal of Time Series Analysis Vol. 8, No. 3, 261-275
[7] Mohammadipour M 2009 Intermittent Demand Forecasting With Integer Autoregressive Moving Average Models Doctoral Dissertation Brunei University
[8] Jung R C and Tremayne A R 2006 Coherent forecasting in integer time series model International Journal of Forecasting 22, 223–38
[9] Freeland R K and McCabe B P M 2004 Analysis of low count time series data by Poisson autoregression Journal of Time Series Analysis 25(5): 701-722
[10] Bu R and McCabe B 2008 Model selection estimation and forecasting in INAR(p) models: A
likelihood-based Markov Chain approach International Journal of Forecasting 24(1): 151-162
[11] Cryer J D and Kung-Sik C 2010 Time series analysis: With application in R (2nd ed.)
[12] Klugman S A and Willmot G E 2019 Loss Models: From Data to Decisions, 5th ed John Wiley & Sons, Inc