PAPER • **OPEN ACCESS**

Generalized Linear Model Approach for Time Series Count Data on Number of Foreign Tourists Modeling in West Java

To cite this article: E Rohaeti *et al* 2021 *J. Phys.: Conf. Ser.* **1863** 012025

Journal of Physics: Conference Series **1863** (2021) 012025 doi:10.1088/1742-6596/1863/1/012025

Generalized Linear Model Approach for Time Series Count Data on Number of Foreign Tourists Modeling in West Java

E Rohaeti[1,2], A Kurnia[2] , K Sadik[2*], and I M Sumertajaya[2]

[1] Department of Mathematics, Pakuan University, Bogor, Indonesia
[2] Department of Statistics, IPB University, Bogor, Indonesia

*E-mail: kusmans@apps.ipb.ac.id

Abstract. This study aimed to model the number of foreign tourist visits in West Java. The number of a tourist visit is part of the discrete data so that the normal distribution approach become less precise in common modeling. In this study, the number of tourist visits was carried out using the Generalized Linear Model approach, combining the Poisson distribution and negative binomial distribution with the identity and log link function. The effect of internal covariates due to a surge in the number of tourists in a given month was also added to the modeling. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. The results showed that the four models obtained were equally good based on the mean absolute percent error (MAPE) values, while the model obtained with the negative binomial distribution integral probability log link function is the best model based on the Akaike information criterion (AIC), Bayesian information criterion (BIC) and integral probability transform (PIT) histogram values. The negative binomial with the log link function approach was then used to model and to predict the number of foreign visits. Plot using negative binomial and log link function, has a value closer to the actual data plot, also strong with the smallest AIC and BIC values.

1. Introduction

Discrete data can only take certain values that are never negatively shaped. In addition to the calculation process of specific individual groups, the calculation of discrete data can also be done at a specific time interval to become a data count time series. Modeling for discrete time series data can be done with the Generalized Linear Model (GLM) approach. Modeling for time-series data using a GLM approach is [8] has developed packages to analyze discrete time series data with the GLM approach. Meanwhile, [1] estimated parameters on the model with a quasi maximum likelihood estimator (QMLE) method.

One of the observations that use discrete time series data is in the tourism field, e.g., tourists' number. According to [3], foreign tourists are every visitor who visits a country outside his residence, driven by one or several purposes without intending to earn in the place visited, and the duration of the visit is not more than 12 (twelve) months.

The number of foreign tourists visited in West Java in the first two months of 2019 increased by 16.4% compared to the same period in 2018. This spike resulting in the number of visitors in a given month is much higher than usual. These spikes result in data values that are much different from others to be identified as remoteness. It is further regarded as a covariate effect of internal influences in the data, called the internal covariate effect. [7], [8].

The foreign tourist data is discrete based on its data type, so this study will be modeled using a GLM approach with Poisson and the negative binomial distribution. The goal is to compare model performance with Poisson and negative binomial distribution approach using identity and log link function, then the best performance AIC and BIC model is used to predict the number of foreign tourists in the future.

2. Literature Review

From [2] and [8], suppose there is discrete time series data Y_t with $t \in \mathbb{N}$. Next, we model the conditional mean $E(Y_t|\mathrm{F}_{t-1})$ from discrete time-series data, for example λ_t and $t \in \mathbb{N}$. Then the general GLM model for modeling discrete time series data is as follows:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-ik}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-jl}) + \eta^T X_t \qquad (1)$$

With : $g: R^+ \to R$ is the link function and $\tilde{g}: N_0 \to R$ is a transformation function, a vector parameter $\eta = (\eta_1, \dots, \eta_r)^T$. In GLM $v_t = g(\lambda_t)$ called the linear predictor, the regression can be used for the past time response variables, defined $P = \{i_1, \dots i_p\}$ and i is integer $0 < i_1 < i_2 \dots < i_p < \infty$, with $p \in N_0$. In the GLM model for discrete time series data, it is possible to regress observed lag $Y_{t-i_1}, Y_{t-i_2}, \dots, Y_{t-i_p}$. The same analogy with lag in observation, defined Q is Q $\{j_1 < j_2 < \cdots < j_p < \infty$ with i is integer and $q \epsilon \mathbb{N}_0$ for the regressor variable on the lag for the conditional mean $\lambda_{t-j_1}, \lambda_{t-j_2}, \dots, \lambda_{t-j_p}$.

The model in equation (1) depends on the link function used. Here is an example of the identity link function, for example, $g(x) = \overline{g}(x) = x$. Then $P = \{1, \dots, p\}$, Q = $Q = \{1, \dots, q\}$ and $\eta = 0$. When $\eta = 0$ then there is no effect of the covariates. The model in equation (1) will then be:

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k (Y_{t-k}) + \sum_{l=1}^{q} \alpha_l (\lambda_{t-l}) \qquad (2)$$

This equation assumes that Y_t has a Poisson distribution.

A further example with a logarithmic link function $g(x) = \log(x), \overline{g}(x) = (x+1)$ and P, Q as defined previously. The equation (1) will be a log-linear model with p and q for discrete time series data analysis. Suppose $v_t = \log(\lambda_t)$ then the equation (1) akan menjadi :

$$v_t = \beta_0 + \sum_{k=1}^{p} \beta_k (Y_{t-k}) + \sum_{l=1}^{q} \alpha_l (v_{t-l}) \qquad (3)$$

The model in equation (1) assumes the Poisson distribution Poisson $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}\,(\lambda_t)$ with the distribution:

$$\mathrm{P}(Y_t|\mathcal{F}_{t-1}) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \qquad y = 0,1 \dots. \qquad (4)$$

This distribution has $\mathrm{var}(Y_t|\mathcal{F}_{t-1}) = \mathrm{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$.

The equation model (1) is the quasi conditional maximum likelihood (ML) estimator. If the observation is assumed to be Poisson distribution, it will be the ordinary ML estimator. For example $\boldsymbol{\theta} = (\beta_0 > 0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q, \eta_1, \dots, \eta_r)^T$ is a vector containing regression parameters. Based on equation 2, if the Data is assumed to be Poisson distribution, the parameters for the equation model (2) will be determined as follows :

$$\Theta = \{\theta \epsilon \mathbb{R}^{p+q+r+1};\ \beta_0 > 0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q, \eta_1, \dots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l \ < 1\}$$

Equation 3 will have a parameter estimator with the following conditions:

$$\Theta = \{\theta \epsilon \mathbb{R}^{p+q+r+1};\ |\beta_1|, \dots, |\beta_p|, |\alpha_1|, \dots, |\alpha_q|, < 1, \left|\sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l\right| \ < 1\}$$

For the observation vector $\boldsymbol{y} = (y_1, \dots, y_n)^T$ a conditional quasi log-likelihood function can be written as follows : [8]

$$\ell(\theta) = \sum_{t=1}^{n} log P_t(y_t; \theta) = \sum_{t}^{n} (y_t \ln(\lambda_t(\theta)) - \lambda_t(\theta)$$

with $p_t(y; \theta) = P(Y_t = y|\mathcal{F}_{t-1})$ is the probability density function of the Poisson distribution Quasi maximum-likelihood estimator (QMLE) $\widehat{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$ assuming that there is a solution to this equation is by optimization:

$$\hat{\theta} = \hat{\theta}_n = \arg max_{\theta \epsilon \Theta}\, \ell(\theta)$$

To assess the goodness of the model has been developed by [9], then by [5] using the integral probability transform (PIT) displayed in the form of a histogram, if the histogram shape is approaching uniform distribution, then the model is better. Besides, [10] another way to assess the model's habit is to use a selection criteria model such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), that is, the model with the smallest AIC and BIC values is the best model.

3. Methodology

3.1. Data
The data used is foreign tourist data in West Java from January 2010 until December 2019. The monthly data of total foreign tourist visits. Data sourced from the Badan Pusat Statistik (BPS) of West Java province. The data used is displayed in Table 1.

Table 1. Number of tourists in West Java January 2010 to December 2019

	Year									
Month	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
January	6678	9669	9737	14077	16397	10453	11065	8614	11600	12529
February	6990	8912	10771	12088	14618	13138	8497	13410	12302	15172
March	7285	9224	13366	16815	21538	15224	15971	17439	15793	16440
April	6984	9949	12711	14068	13631	16978	30922	16189	14248	14830
May	8358	9592	12829	18023	14725	18902	16841	15452	10571	8168
June	7868	11262	15533	16640	16942	15423	9055	8360	6493	8881
July	8531	12020	11736	7803	6241	6688	9499	12017	12814	12645
August	7408	6673	7194	8808	10648	10409	12663	15189	13766	14129
September	5410	7138	13749	14742	14132	10652	15141	14120	13399	13028
October	9799	9281	7537	11984	15086	10755	17444	14151	13918	13569
November	6598	11265	15017	18243	16644	14951	12876	15541	14364	14715
December	10570	12565	18265	24401	20880	17067	22410	18031	17375	15159
Total	92479	117550	148445	177692	181482	160640	182384	168513	156643	159265

3.2. Stages of analysis
Stages of analysis in the model as follow:

a. Explore data to see a General data overview through a plot of the number of foreign tourist arrivals in West Java Province from 2010 to 2019.
b. Define the P and Q are initial values by forming an ACF and PACF plot, specifying the optimal combination based on the smallest AIC value.
c. Modeling with the Poisson and the negative binomial distribution approach with identity (normal) and log link functions, adding internal covariate effects, and Estimating model parameters with the Quasi maximum-likelihood estimator (QMLE).
d. Comparing the performance of models with the Poisson and Binomial negatives distribution in modeling data on the number of foreign tourists visiting West Java, the best model was obtained.
e. Use the best model of the results gained at Stage 5 to predict the number of tourist visits
The analysis was conducted using the R 3.6.0 software with the package Tscount with the tscount function.

4. Result and Discussion

4.1. Data Exploration

Before modeling, analyze the number of foreign tourist arrivals in West Java province from 2010 to 2019. An overview of the data patterns of monthly foreign tourists visiting presented in Figure 1. Figure 1 shows that Data on the number of foreign tourist's visits to West Java is relatively stationary as it fluctuates around the mean value, although in some of the last data, it tends to form a downward trend. Figure 1 also looks at some values bursting with the data 40 and 84, which is considered an intercept less. It is further regarded as a covariate effect of internal influences in the data, called the internal covariate effect. The Covariate effect was incorporated in the modeling process based on previous time observation and the previous time conditional mean [8] so that the data will be added a coefficient is resulting from the intervention.

Theoretically, if using the data type which is discrete Data, then modeling with the generalized linear approach of the model is done by the Poisson distribution approach and the log link function, but in this research will be attempted with identity (normal) hyphen function in addition to the log Link function, then corrected using a negative binomial distribution with identity and log link function, next a tested, the best approach model is used to predict the number of tourist visits in the future.

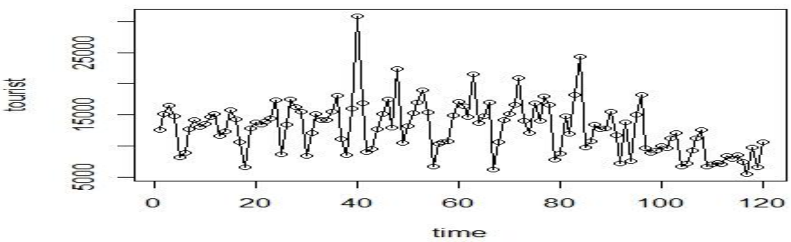
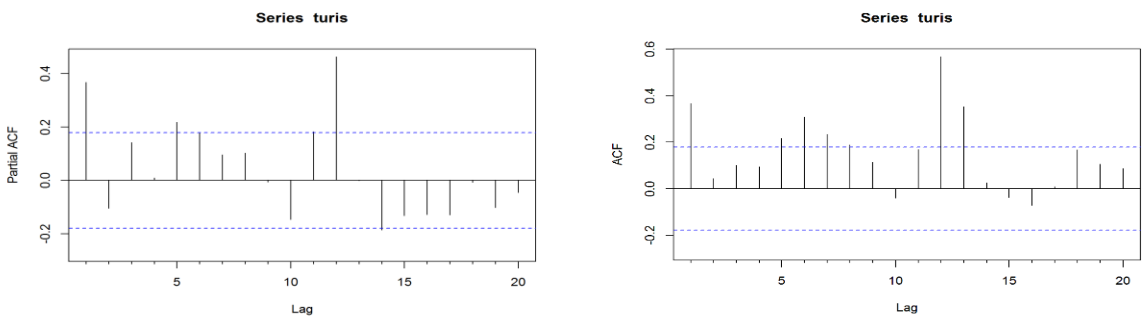


Figure 1. A plot between foreign tourist data and the time

4.2. P and Q Value Determination

Before modeling the initial step, determine P and Q's value based on the autocorrelated function (PACF) and autocorrelated function (ACF) plot. The result shows in Figure 2.



(a) (b)

Figure 2. (a) ACF plot and (b) PACF plot

Figure 2 shows that the PACF line crosses the 1.96 limits at the 1.5 lag and 12, while the ACF line crosses the 1.96 limits on the 1.5 to-1, 6, 7, 12, and 13. So that the value of P and Q to be tested is P = {1, 5, 12} and Q = {1, 5, 6, 7, 12, 13}, then from the value P and Q is determined the optimal combination based on the smallest AIC value. [1]. The following is presented a combination of P and Q based on the Poisson distribution approach :

Table 2. AIC value model with negative binomial distribution and identity link function

P,Q	1	5	6	7	12	13
1	157098.1	157098.2	155985.1	157098.2	157098.1	155666.5
5	157193.8	157141.0	157123.1	157298.6	157231.5	157105.1
12	156217.1	153254.8	153225.1	**153064.6**	156217.1	156217.1

Of the several combinations listed in Table 2, the optimal combination is P = 12 and Q = 7, and then these values are used in modeling.

4.3. Modeling with Poisson distribution and identity link functions

Once obtained, P and Q's value, namely P = 12 and Q = 7 based on the combination result with the smallest AIC value. Then done modeling for the number of foreign tourists visit in West Java by using packages Tscout [8] and identity (normal) Link function. Modeling begins with the estimation of the model parameters, and the result of the estimation of parameter models can be seen in Table 3.

Table 3. Model parameter interview with Poisson distribution and identity link function

Parameter	Estimate	Std.Error	CI(lower)	CI(upper)
β_0	1.28×10^4	8.64×10^2	1.11×10^4	1.45×10^4
β_{12}	4.59×10^{-2}	2.91×10^{-3}	4.02×10^{-2}	5.16×10^{-2}
α_7	1.04×10^{-10}	6.45×10^{-2}	-1.26×10^{-1}	1.26×0^{-1}
Interv_1	1.25×10^{-1}	2.28×10^1	-4.45×10^1	4.47×0^1
Interv_2	6.78	1.20×10^2	-2.29×10^2	2.42×10^2

According to Table 3, foreign tourists visit model in West Java with a GLM approach and Poisson distribution as follows :

$$\hat{\lambda}_t = 1.28 \times 10^4 + 4.59 \times 10^{-2} Y_{t-12} + 1.04 \times 10^{-10} \lambda_{t-7} + 0.125(t = 40) + 6.78(t = 84)$$

With time 40 and 84 as internal covariate effects due to the data, there was a surge in foreign tourist arrivals. This model indicates that the average of foreign tourists visit in year *t* is influenced by the number of tourists visited in the previous 12 months or a year in advance. Average visitors also influence it in the previous seven months. Since there is a variable intervention than in certain months, a year of intervention will be added with a specific coefficient as it is written on the model.

4.4. Modeling with Negative Binomial distribution and identity link functions

Further modeling is carried out with a negative binomial distribution approach and uses the identity link function, with the following same stages presented a combination of P and Q values to determine the optimal combination based on the smallest AIC value, as follows:

Table 4. AIC value model with negative binomial distribution and identity link function

P,Q	1	5	6	7	12	13
1	2341.650	2341.650	2340.737	2341.650	2341.650	2340.537
5	2341.591	2341.640	2341.592	2341.792	2341.728	2341.603
12	2340.956	2338.936	2338.958	**2338.736**	2340.956	2340.956

Table 4 shows that P and Q's value for the negative binomial distribution is still the same as the Poisson distribution at P = 12 and Q = 7. However, the resulting AIC value is smaller than the AIC value of the Poisson distribution. After this, the estimation of the model parameter with a negative binomial distribution, as the result of the alleged parameters as follows:

Table 5. The parameter with negative binomial distribution and identity link function

Parameter	Estimate	Std.Error	CI(lower)	CI(upper)
β_0	1.28×10^4	3.10×10^4	-47890.78	7.36×10^4
β_{12}	4.59×10^{-2}	1.05×10^{-1}	-0.16	2.52×10^{-1}
α_7	1.04×10^{-10}	2.31	-4.53	4.53
Interv_1	1.25×10^{-1}	8.18×10^2	-1602.55	1.60×10^3
Interv_2	6.78	4.38×10^3	$-8.568.37$	8.58×10^3
ϕ	10.43			

Table 5 shows the result of suspected parameters equal to the alleged Poisson distribution. However, the resulting standard of error is different. In addition to the case of overdispersion, a negative binomial distribution approach can suspect the magnitude of the overdispersion parameter, i.e., $\phi = 10.43$. Based on the model for the number of foreign tourists visit in West Java with the Negative Binomial distribution given by $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}\,(\lambda_t, \phi)$:

$$\hat{\lambda}_t = 1.28 \times 10^4 + 4.59 \times 10^{-2} Y_{t-12} + 1.04 \times 10^{-10} \lambda_{t-7} + 0.125\,(t = 40) + 6.78(t = 84)$$

After modeling with the normal approach or identity, the goodness of models produced based on the AIC value obtained is still not good; subsequent modeling is done with the log link function.

4.5. Poisson distribution assumption Modelling and log link function

The modeling stage with the Poisson distribution and the log link function approach is the same as the previous modeling stage. Table 6 presents a preliminary step of modeling that determines the optimal combination of P and Q values as follows:

Table 6. AIC value models with Poisson distribution with the Log link function

P, Q	1	5	6	7	12	13
1	122295.81	119128.68	111308.50	124773.59	106418.17	120266.77
5	123179.68	135936.45	133459.20	112329.03	124433.34	134104.01
12	**75761.79**	85234.16	84115.66	83862.21	84902.41	82677.64

The optimal combination of Table 6 is obtained based on the smallest AIC value, which is located at P = 12 and Q = 1, and after this, the value is used in modeling with the alleged model parameters obtained in Table 7 below:

Table 7. Model parameter interview with Poisson distribution and log link function

Parameter	Estimate	Std.Error	CI(lower)	CI(upper)
β_0	0.000674	0.033529	-0.0590	0.0725
β_{12}	0.67654	0.002912	0.6708	0.6822
α_1	0.32275	0.003483	0.3159	0.3296
Interv_1	-0.04096	0.000731	-0.0424	-0.0395
Interv_2	0.20936	0.006743	0.1961	0.2226

4.6. Modeling assumption of the negative binomial distribution and log link function

As the AIC value of Poisson's approach is still considerable and cannot suspect the overdispersion parameters, then modeling with the negative binomial approach and the log link function, tapping the stage carried out is the first determining the optimal combination of P and Q based on the smallest AIC value, as well as a result in Table 8 follows:

Table 8. AIC models with a negative binomial distribution and log link function

P, Q	1	5	6	7	12	13
1	2312.475	2306.794	2297.668	2313.334	2291.161	2309.469
5	2310.155	2326.519	2323.607	2300.273	2311.444	2322.995
12	**2249.530**	2265.730	2263.021	2263.539	2264.759	2260.396

Once the optimal combination of P = 12 and Q = 1, then the value is used in the modeling. After this, the estimate with the result of the alleged parameters presented in Table 9 as follows:

Table 9. The parameter with negative binomial distribution and log link function

Parameter	Estimate	Std.Error	CI(lower)	CI(upper)
β_0	0.00674	0.8628	-1.6843	1.69777
β_{12}	0.67654	0.0778	0.5240	0.82904
α_1	0.32275	0.0895	0.1474	0.49811
$Interv_1$	-0.04096	0.0191	-0.0784	-0.00355
$Interv_2$	0.20936	0.2279	-0.2373	0.65603
ϕ	18.68			

4.7. Comparison of model performance

The next stage evaluates the four approaches' model performance based on the PIT histogram criteria presented by [8]. The Histogram PIT for Poisson and negative binomial distribution with the identity (normal) Link function shows as follows:

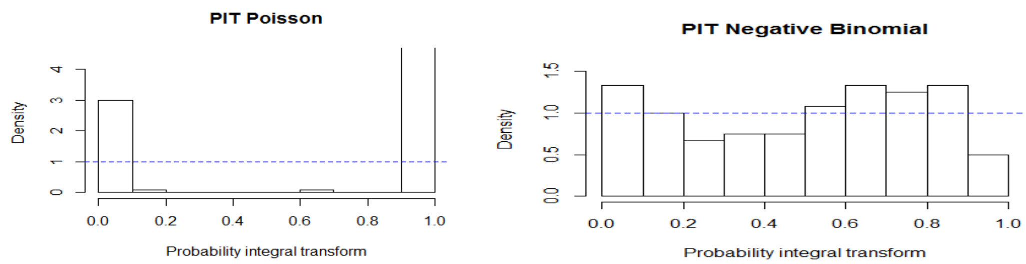


Figure 4. Histogram PIT with the identity link function

Figure 4 shows that the PIT histogram form of the model with Poisson distribution is still far from the uniform distribution form, so it can be said that the resulting model is still not acceptable [6].

Subsequent performance evaluations based on PIT histogram criteria were performed on models with the Poisson distribution approach and negative binomial with the log link function, and the results are presented in Figure 5.

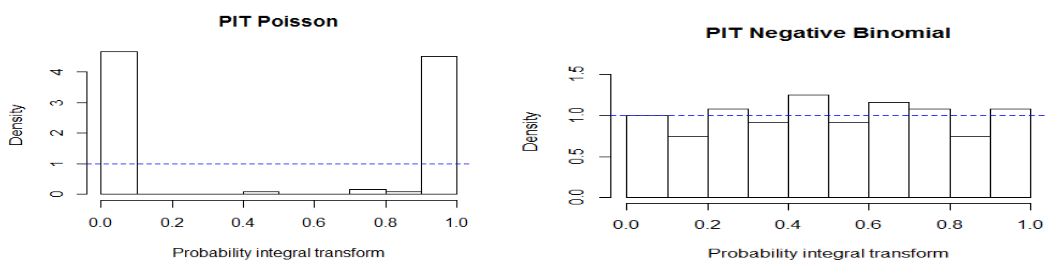


Figure 5. Histogram PIT with a log link function

Figure 5 shows that the Poisson distribution model and the Link log function are still not good because it has not approached the uniform distribution form. In contrast, a model with a negative binomial distribution with the log link function has approached the uniform; the histogram form is better than the identity-link function. It means that the model with a negative binomial approach with the Log Link function is most well-performing compared to the other approach. Comparison of performance models if based on MAPE, AIC, BIC, and histogram PIT values are presented in the following table:

Table 10. Comparison of model performance

Model	MAPE	AIC	BIC	Histogram PIT
Poisson identity	18.413	153225.1	153239.1	Not close to uniform distribution
Negative Binomial identity	18.413	2338.958	2355.683	Close to uniform distribution
Poisson log	18.060	75761.79	75775.72	Not close to uniform distribution
Negative Binomial log	18.00	2249.53	2266.255	Close to uniform distribution

In Table 10, the model with the same link function resulted in the same MAPE that showed the model as well, but if it is seen from the AIC, BIC, and the PIT histogram values that the model with a negative binomial distribution is the best model among the four model approaches so that the model with a negative binomial. The goodness of the model is also demonstrated by the actual plot of data comparison with modeling results using the following four approaches:

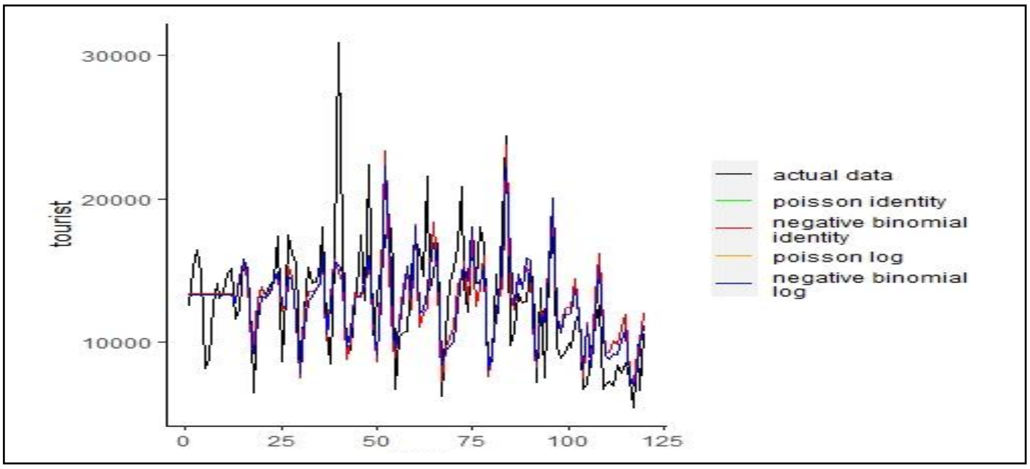


Figure 6. Comparison of actual data and models

Figure 5 shows the data plot with the other four estimators. This plot shows that all the plot estimates have a reasonably good value because they follow actual data pattern data. However, the blue plot, namely the estimation with negative binomial and log link function, has a value closer to the actual data plot, also substantial with the smallest AIC and BIC values.

4.8 Forecasting the number of foreign tourists

Based on the performance comparison of the models that have been produced, the best models are used to predict the next year (January-December 2020). The forecasting results are presented in the following Table 11. From this result, we can see that there are fluctuations in the forecast of foreign tourist arrivals; however, as in the original data, previously, foreign visitors mostly occur during December.

Table 11. Forecasting results

Month	Y	Month	Y
January	7874	July	8322
February	7268	August	7695
March	7283	September	6066
April	7083	October	8396
May	7926	November	7136
June	7890	December	9313

5. Conclusion
Based on MAPE, the models have the same of the goodness of fit, but if based on the additional criteria for the goodness of fit model from the AIC, BIC, and PIT histogram values, we can see that model with the negative binomial distribution approach and the log link function is better than another model. This model is also suitable for forecasting foreign tourists in West Java in the future. The results show that the seasonal factor has a positive effect, which means that any increase in the number of tourists in the previous year will cause an increase in the number of tourists in the current month, and the expected value in the previous month also has a positive effect on the expected value of visitors at present.

References
[1] Ahmad A, Francq C 2016 Poisson QMLE of Count Time Series Models *Journal of Time Series Analysis*, 37(3):291–314. doi:10.1111/jtsa.12167.
[2] Bosowski N, Manolakis D 2017 Generalized Linear Models for Count Times Series *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). doi:10.1109/ICASSP.2017.7952962.
[3] BPS. 2020. Concept and Definition of International Tourist. Jakarta. www.bps.go.id
[4] Christou V, Fakianos K 2014 Quasi Likelihood Inference for Negative Binomial Time Series Model *Journal Of Time Series Analysis*. 35(1):55-78. doi:10.1111/jtsa.12050.
[5] Christou V, Fakianos K 2015 Estimating and Testing Linearity for Non Linear Mixed Poisson Autoregression *Electronic Journal Of Statistic*. 9(1):1357–1377.
[6] Fokianos K, Fried R 2010 Interventions In INGRACH Processes *Journal of Time Series Analysis*. 31.210-225.doi:10.1111/j.1467-9892.2010.00657.x.
[7] Fokianos K, Fried R 2012 Interventions in Log-Linear Poisson Autoregression *Statistical Modelling*. 12(4): 299–322.
[8] Liboshick T, Fakianos K, Fried R 2017 tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models *Journal of Statistical Software*. 82(5). DOI: 10.18637jss.v082.i05.
[9] Jung R, Tremayne A 2011 Useful Models for Time Series of Counts or Simply Wrong Ones? *AStA Advances in Statistical Analysis*. 95:(1).59–91.doi:10.1007/s10182-010-0139-9.
[10] Rochyati I, Syafitri UD, Sumertajaya IM 2019 *Study of Forecasting Models on Foreign Tourists Visit and International Passengers Arrival without and with Covariates*. [thesis]. IPB University.