Applied Mathematics and Computational Intelligence
Volume 11, No.1, Dec 2022 [104-113]

Comparing Model of Air Pollution Index Using Generalized Autoregressive Conditional Heteroskedasticity Family (GARCH)

Nurul Asyikin Zamrus[1], Mohd Hirzie Mohd Rodzhan[2]* and Nurul Najihah Mohamad[3]

[1,2,3] Department of Computational and Theoretical Sciences, Kulliyyah of Science, International Islamic University Malaysia 25200 Kuantan, Pahang, Malaysia

[2*] Corresponding author: mohdhirzie@iium.edu.my

Received: 25 October 2021; Accepted: 11 February 2022; Available online (In press): 16 March 2022

ABSTRACT

The Air Pollution Index (API) of Malaysia has increased consistently in recent decades, becoming a serious environmental issue concern. In this paper, the daily integer value time series data for API in Penang and Sarawak from January to June in 2019 using generalized autoregressive conditional heteroskedasticity (GARCH) family for discrete case namely Poisson integer value GARCH (INGARCH), negative binomial integer value GARCH (NBINGARCH) and integer value autoregressive conditional heteroskedasticity (INARCH) models are analysed. The parameters of the models will be estimated using quasi likelihood estimator (QLE) and compared their Akaike information criterion (AIC) to determine the best model fitted the data. The results showed that INGARCH (1,1) model will be the best model because it has the small value of AIC. Hence, the findings are very important for controlling the API results in the future and taking protective measures for the conservation of the air.

Keywords: Time series, Generalized Autoregressive Conditional Heteroskedasticity (GARCH), Air Pollution Index, Integer-Value.

1 INTRODUCTION

Air quality prediction has become a crucial area of environmental science due to the negative effects of high concentrations of different contaminants on human health. The air quality also can be defined as API. Air pollution consists of a combination of gases and particulate matter in a negative way quantities released into the atmosphere by natural or human-made events [1]. Carbon monoxide (CO), ozone (O3), particulate matter (PM10), nitric oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), and sulphur dioxide (SO_2) are the gases and particles involved. Therefore, air pollution is a big issue in many parts of the world, and it poses two major issues. . First, there's the influence on human health, and then there's the impact on the environment [2]. It is exceedingly dangerous and has resulted in several deaths around the world. Air pollution was estimated to be responsible for 200,000 to 570,000 deaths per year worldwide, or around 0.4 to 1.1 percent of all deaths [3].

However, time series analysis has been used by many researchers in the literature to predict the concentration of various pollutants and the air quality [4]. The data of eight stations in central Taiwan have been analysed by using multivariate time series analysis models namely Autoregressive

Conditional Heteroskedasticity (ARCH) and GARCH [5]. The models selected both the photochemical and fuel factors for evaluating the various time series patterns. A hybrid model is proposed to deal with both linear and nonlinear data of a station in Delhi from 1999 to 2003 [6]. The performance of the linear, nonlinear and hybrid model was checked using mean absolute percentage error and relative error. It was found that the hybrid model outperformed both the linear and nonlinear models. Next, there is researcher analysed the patterns of the relationship between various air pollutants of an Alpine Italian province [7]. The dynamic multiple time series analysis is carried out using a common autoregressive stochastic model to find the improvement level in pollution during the last decade. Based on the Kalman filter-based autoregressive model, Hoi and Mok [8] introduced the time-varying autoregressive model with linear exogenous input (TVAREX) for predicting daily PM10 concentrations. . The results of the TVAREX model were compared to the artificial neural network (ANN) model and it was observed that TVAREX outperformed ANN. Kadiyala et al. [8] developed a model to manage indoor air quality using a multivariate time series model to manage the concentrations of both carbon dioxide and carbon monoxide. This prediction was applied to design an optimal ventilation system for vehicles. Kumar et al. [9] predicted the air quality index of Delhi based on three models namely Autoregressive Integrated Moving Average (ARIMA), principal component regression and hybrid of the first two. It was found that model (3) demonstrated the highest performance accuracy compared to other models. Further, the importance of various meteorological parameters in model 3 was assessed based on principal component analysis. The short-term prediction of the concentration of ozone in Albany, New York was presented by Tsakiri et al. [10] based on the vector autoregressive model and the Kalman filter. The prediction was found to be most accurate when the time series components of temperature and solar radiation were taken into consideration.

Though the majority of the research has been focused on the prediction of individual concentrations of pollutants, there is a need to predict a single value that indicates the air quality. In this paper, the univariate time series analysis of the Samarahan, Limbang, Seberang Perai and Seberang Jaya has been performed by using the GARCH family namely INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) models. The detail about study area regarding the data description has been presented in the next section. Section 3 discusses the methodology for time series analysis of the air index pollution dataset. A comparison between the performance evaluation results between INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) have been discussed in Section 4. Hence, a conclusion has been presented in the last section.

2 MATERIAL AND METHODS

In this section the data description and research methodology will be discussed.

2.1 Data description

The dataset used in this research is API which is retrieved from the Department of Statistic Malaysia open-source data website [11]. The frequency of the data collection is daily data within six (6) months from January 2019 until June 2019. The data have been chosen from eight (8) locations which are located at Penang and Sarawak. Penang and Sarawak were chosen for this study because of the mix of urban and suburban sites available. Furthermore, transboundary contamination from neighbouring nations has frequently impacted these areas, which has been the primary cause of hazardous events. This dangerous haze is caused by forest fires in Sumatra and Kalimantan. Due to

the haze that hit the country, Malaysia ranks third in the world in the list of countries that record the highest API after Iran and Indonesia in 2019. In 2019, based on the API observations by the World Air Quality Index (WAQI) Malaysia recorded an API reading of 271 while Iran and Indonesia are 385 and 303 respectively. In total, six areas still recorded very unhealthily. API readings including Kuching, Samarahan, Sri Aman, Sibu and Sarikei in Sarawak and Balik Pulau in Penang. Sarawak and Penang give a big impact on this air pollution because the location is near to the Sumatra and Kalimantan where the forest fire happens. This makes the spread of the haze is more impactful in this area due to the direction of the wind blow faster the process of air pollution to happen here. Hence, due to the haze, a total of 2,649 schools were closed including in Penang and Sarawak while the number of asthma and conjunctivitis cases was found to be increasing based on monitoring from 31 haze sentinel clinics. Hence, this area has been choosing in this research to measure the air quality using the GARCH family model.

2.2 Methodology

Time series analysis using INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) models have been carried out for the prediction of the air pollution index. The steps of the methodology of the time series analysis using INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) model have been summarized in the subsection below:

2.2.1 Comparing GARCH Family Analysis

The INGARCH, NBINGARCH and INARCH models are fitted in the form of (1,1) by QLE. The Poisson assumption is right to get a standard ML estimator. However, if we assume a mixed Poisson distribution, we get a quasi-ML estimator. The vector of regression parameters is denoted by the symbol $\theta = (\beta_0, \beta_1, \ldots \beta_p, \alpha_1, \ldots, \alpha_q)$. The parameter space for the INGARCH (1,1) model with covariates is given regardless of the distributional assumption [13].

$$\Theta = \left\{\theta \in R^{p+q+r+1};\ \beta_0 > 0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1\right\} \quad (1)$$

The intercept β_0 must be positive and all other parameters must be nonnegative to ensure the positivity of the conditional mean. The other condition ensures that the fitted model has a stationary and ergodic solution. Next, the efficacy of the ARCH is investigated. Until running simulations with the time series combination of ARCH and GARCH models, the model calibration phase must be completed first to ensure that the residual series is not connected to the first-order series, often known as white noise and that the model is acceptable. In testing the presence of the ARCH effect, the generalised autoregressive representation of the squared residuals ($\hat{u}_t^2$) with the error (e_t) is given as:

$$\hat{u}_t^2 = b_0 + b_1 u_{t-1}^2 + b_2 u_{t-2}^2 + b_3 u_{t-3}^2 + \ldots + b_q u_{t-q}^2 + e_t \quad (2)$$

The Lagrange multiplier (LM) test is being used to test for the arch effect before running the GARCH family model. Consider the null hypothesis (H_0) of no ARCH errors versus the alternative hypothesis (H_1) that the conditional error variance is given by an ARCH (q) process. The null hypothesis will be rejected if p-value less than 0.05. Iterative nonlinear calculations for estimating model parameters can only be done with the model that has ARCH effectiveness [13].

Then, the data were tested for stationarity using the Augmented Dickey–Fuller (ADF) test, with the H_0 being that the time series is non-stationary and the H_1 state that the time series is stationary. The ADF test revealed that when p-value less than 0.05, indicates that H_0 is rejected and H_1 is accepted which means the result is significant. The INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) model's values were chosen based on the (AIC) value, which is given by (3).

$$AIC_{p,q} = \frac{-2ln(maximized\ likelihood)+2r}{n} \approx ln(\sigma_i^2) + r\frac{2}{n} + constant \quad (3)$$

where n is the number of data observations, $r = p + q + 1$ and σ_i^2 is the maximum likelihood prediction.

Table 1 shows the AIC values for different α and β parameters. The different values of α and β parameters were tested ranging from 0 to 5, while the order (p, q) for the model were chosen to be 1 based on the ADF test. Hence, given that the best INGARCH (1,1) model have the lowest AIC values is (1,1).

2.2.2 Evaluation Model Performance

It's critical to evaluate model performance in order to find the optimal model with the least amount of error. There are a number of statistical tests that may be used to assess model validity. In most cases, the AIC is used to assess modelling accuracy. The AIC for the GARCH family model is then calculated as follows:

$$AIC = 2(p + q + 1)2lnL \quad (4)$$

where lnL denotes the log-likelihood function, T is the number of non-missing data, p denotes the ARCH component model order, and q denotes the GARCH component model order.

In this paper, three models under the GARCH family have been chosen which are called as INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) model.

Table 1: Performance of the proposed model.

Location	Parameter	INGARCH (1,1)	NBINGARCH (1,1)	INARCH (1,0)
Seberang Perai	μ	17.30	17.30	17.31
	α	5.76^{-8}	5.76^{-8}	-
	β	0.68	0.68	0.68
	AIC	1199.53	1199.53	1197.53
	BIC	1209.10	1209.10	1203.91
Balik Pulau	μ	12.59	12.59	14.85
	α	0.09	0.09	-
	β	0.68	0.68	0.72
	AIC	1209.55	1209.55	1208.09
	BIC	1219.11	1219.11	1214.46
Kimanis	μ	4.73	4.73	5.29
	α	1.77^{-5}	1.77^{-5}	-

	β	0.89	0.89	0.88
	AIC	1252.21	1243.73	1250.88
	BIC	1261.77	1256.48	1257.26
Limbang	μ	11.4	11.4	11.14
	α	2.76^{-5}	2.76^{-5}	-
	β	0.714	0.714	0.721
	AIC	1399.071	1324.02	1397.10
	BIC	1408.63	1336.77	1403.475
ILP Miri	μ	10.62	10.60	11.62
	α	0.01	0.01	-
	β	0.77	0.77	0.76
	AIC	1207.138	1205.01	1203.231
	BIC	1214.57	1219.89	1209.61
Bintulu	μ	15.30	15.30	15.29
	α	6.95^{-4}	6.95^{-4}	-
	β	0.70	0.70	0.70
	AIC	1243.47	1243.47	1240.88
	BIC	1253.03	1253.03	1247.25
Kapit	μ	9.73	9.73	15.80
	α	0.25	0.25	-
	β	0.47	0.47	0.56
	AIC	1220.52	1210.81	1224.38
	BIC	1230.08	1223.56	1230.76
Samarahan	μ	6.59	6.59	11.92
	α	0.30	0.30	-
	β	0.50	0.50	0.64
	AIC	1286.38	1251.58	1292.46
	BIC	1295.94	1264.33	1298.83

2.2.2.1 INGARCH (1,1)

An INGARCH model is proposed based on Ferland et al., [12], which is defined as follows:

$$\lambda_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} \quad (5)$$

where $\gamma_0 > 0, \gamma_i \geq 0, 1 \leq i \leq p, \beta_j \geq 0, 1 \leq j \leq q$.

2.2.2.2 NBINGARCH (1,1)

Let $\{X_t\}$ be a time series of counts data. The random variables $X_1, \ldots, X_n$ are assumed to be independent conditional on F_{t-1} and the conditional distribution of X_t is given by a normal binomial distribution. To be specific, we consider the following model:

$$X_t | F_{t-1} : NB(r, p_t) \quad (6)$$

where F_{t-1} is the r-field generated by $\{X_t, X_{t-2}, \dots\}$, r is a positive number and p_t satisfies the model

$$\frac{1-p_t}{p_t} = \lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} \quad (7)$$

where $\alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0, i = 1, \dots, p, j = 1, \dots, q, p \geq 1, q \geq 0$.

2.2.2.3 INARCH (1,0)

Weiß [14] refers to the purely autoregressive INARCH (p,0) model as a (p,0) model. The purely autoregressive INARCH $(p, 0)$ model is also called an $(p, 0)$ model by Weiß [14]. The $(p, 0)$ model is defined as follows:

$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \quad \beta_i X_{t-i} \quad (8)$$

where $t \in Z, \beta_0 > 0, \beta_i \geq 0$, and $i = 1, \dots, p$.

3 RESULTS AND DISCUSSION

The time series graph shown in Figure 1 explained that the pattern for the graph is irregular. The following time series plot shows a drastic shift in the Air Pollution Index value after 3 months for all stations. These is due to the begin of haze being spread. These data show random variation which means there are no patterns or cycles. There is high volatility in the graph throughout the dataset from January 2019 until June 2019. The volatility also indicates there is an ARCH effect in our result.

Based on Table 1, the higher mean of API is Seberang Perai with value 53.91. Meanwhile, Samarahan has the lowest mean with value 33.3. The skewness result of Air Index Pollution for Balik Pulau, Limbang and ILP Miri show that the data is substantially skewed distribution because the number is less than -1. Meanwhile, Seberang Perai, Kimanis, Bintulu, Kapit and Samarahan are between -1 and +1 indicates non substantially skewed distribution. The kurtosis for all the datasets is not approached to 3, so the dataset has no “heavy tails” and no “light-tailed”. As a result, distributions with skewness and/or kurtosis that exceed these limits are classified as nonnormal [15]. Hence, for the GARCH family it follows nonnormal behavior indicates that the distribution and result suite our model.

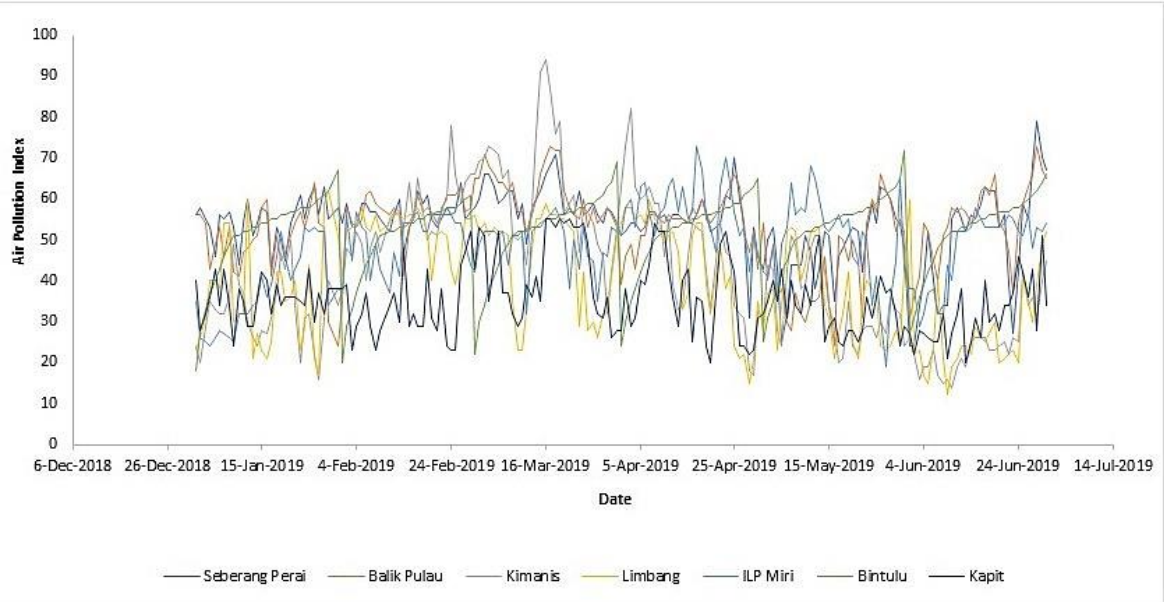

Figure 1: Time Series Graph

The autocorrelation functions (ACF) of time series can be used to infer their stability or instability, as well as memory characteristics. Short-memory processes with non-zero autocorrelations at only a few lags are stable, whereas long-memory processes with major autocorrelations on many lags are unstable. Therefore, in this research lag (1,1) is using for INGARCH and NBINGARCH whereas the INARCH model is using lag (1,0). Stationarity means that the statistical characteristics of a process under study do not change over time [16]. The Air Pollution Index is stationary which means the p-value of the ADF test is less than 0.05. When the result is less than 0.05, the H_0 is rejected and H_1 is accepted. Consequently, the result is significant for the stationary test using the ADF test.

Table 2: Statistical Summary.

Location	Mean	Skewness	Kurtosis	Unit Root Test	Arch Effect
Seberang Perai	53.91	-1.09	2.13	-4.219***	83.0290***
Balik Pulau	53.27	-0.76	0.46	-3.8425***	64.3920***
Kimanis	42.47	0.51	-0.35	-3.6245**	75.3820***
Limbang	39.77	-0.15	-1.41	-3.7054**	48.9190**
ILP Miri	48.53	-0.56	-0.03	-3.4451**	44.5040**
Bintulu	51.75	-1.27	1.42	-5.0867**	44.0160***
Kapit	35.51	0.54	-0.42	-3.8292**	111.0600***
Samarahan	33.3	0.32	-0.75	-3.6815**	64.6250***

NOTE: $^{**}p < 0.05$, $^{***}p < 0.01$

Then, the volatility of the data is being evaluated to make sure there is an ARCH effect in the API database. The ARCH effect is being evaluated by using the LM test. The result shows, there is an ARCH effect in the model because the p-value is less than 0.05. In univariate time series models, the LM test for ARCH is widely used as a specification test [17]. It is an ARCH model against which no conditional

heteroskedasticity is tested. It is a test of no conditional heteroskedasticity against an ARCH model. The LM statistic is asymptotically distributed as X^2 under the null hypothesis.

The performance evaluation results of the model based on Air Pollution Index value only. The data were examined using ADF to evaluate the models' performance. This is to ensure the data is stationary before running the GARCH family model. Hence, the performance for INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) are being compared by using AIC and BIC. The lowest AIC and BIC indicate the best model. The best model for Seberang Perai, Balik Pulau, ILP Miri and Bintulu is INARCH (1,0) model. Furthermore, INGARCH (1,1) model proves that it shows the best model for Kimanis, Limbang, Kapit and Samarahan. By comparing the value of BIC and AIC, it shows that INGARCH (1,1) model is in good agreement with the majority station that has the lowest AIC and BIC for API.

4 CONCLUSION

This paper discussed the models' comparison, namely INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0). The primary purpose of this research was to find the most effective time series approaches in an air quality forecasting model employing the GARCH family model for the daily API in eight locations in Malaysia's Penang and Sarawak. According to the findings of this investigation, the INGARCH (1,1) model was capable of modelling and forecasting API index values. INGARCH (1,1) has proven to be a versatile and intelligent forecasting method that may be used to model complex and poorly understood processes. Compared to the standard ARCH family INARCH (1,0) and NBINGARCH (1,1), the INGARCH (1,1) will be able to produce more accurate predictions for the observed API at all eight sites with a univariate model, where the input came from the best both INARCH (1,0) and INGARCH (1,1) lags (1,1). As a result, we recommend that the simplest INGARCH (1,1) be utilised for future air pollution forecasting for univariate integer values because it is good at predicting fluctuation series with trend and seasonality, such as air quality data.

ACKNOWLEDGEMENT

The authors are grateful to the Malaysian Department of Environment for supplying the air pollution data that enabled them to write this paper.

REFERENCES

[1] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental Pollution*, vol. 151, no. 2, pp. 362–357, 2008.

[2] A. Kurt and A. B. Oktay, "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986–7992, 2010.

[3] World Resources Institute, "World Resources Institute: (2002) Rising Energy Use: Health Effects of Air Pollution," *World Resources Institute,* 2002. [Online]. Available: *http://www.airimpacts.org (2002).*, vol. Accessed J, 2002.

[4] X. Y. Ni, H. Huang, and W. P. Du, "Relevance analysis and short-term prediction of PM 2.5 concentration in Beijing based on multisource data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.

[5] E. M. Y. Wu, S. L. Kuo, and W. C. Liu, "Generalized autoregressive conditional heteroskedastic model for water quality analyses and time series investigation in reservoir watersheds," *Environmental Engineering Science*, vol. 29, no. 4, pp. 227–237, 2012, doi: 10.1089/ees.2011.0086.

[6] A. B. Chelani and S. Devotta, "Air quality forecasting using a hybrid autoregressive and nonlinear model," *Atmospheric Environment*, vol. 40, no. 10, pp. 1774–1780, 2006.

[7] P. Giuliana and M. Paola, "Local atmospheric pollution evolution through time series analysis," *Journal of Mathematics and Statistical Science*, pp. 781–788, 2016.

[8] A. Kadiyala and A. Kumar, "Multivariate time series models for prediction of air quality inside a public transportation bus using available software," *Environmental Progress & Sustainable Energy*, vol. 33, no. 2, pp. 337–341, 2014.

[9] A. Kumar and P. Goyal, "Forecasting of daily air quality index in Delhi," *Science of the Total Environment*, vol. 409, no. 24, pp. 5517–5523, 2011.

[10] K. G. Tsakiri and I. G. Zurbenko, "Prediction of ozone concentrations using atmospheric variables," *Air Quality, Atmosphere & Health*, vol. 4, no. 2, pp. 111–120, 2011.

[11] Department of Statistics Malaysia, "Annual Minimum And Maximum Air Pollutant Index For Selected Stations, Malaysia," *Department of Statistics Malaysia*, 2019. [Online]. Available: https://www.data.gov.my/data/ms_MY/dataset/annual-minimum-and-maximum-air-pollutant-index-for-selected-stations-malaysia

[12] R. Ferland, A. Latour, and D. Oraichi, "Integer-valued GARCH process," *Journal of Time Series Analysis*, vol. 27, no. 6, pp. 923–942, 2006.

[13] E. M. Y. Wu, S. L. Kuo, and W. C. Liu, "Generalized autoregressive conditional heteroskedastic model for water quality analyses and time series investigation in reservoir watersheds," *Environmental Engineering Science*, vol. 29, pp. 227–237, 2012, doi: 10.1089/ees.2011.0086.

[14] C. H. Weiß, "Modelling time series of counts with overdispersion," *Statistical Methods and Application*, vol. 18, pp. 507–519, 2009.

[15] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Cham: Springer International Publishing Switzerland, 2016.

[16] T. Stadnitski, "Time series analyses with psychometric data," *PLoS ONE*, vol. 15, no. 4, pp. 1–12, 2020, doi: 10.1371/journal.pone.0231785.

[17] A. S. Nastić, M. M. Ristić, and A. D. Janjić, "A mixed thinning based geometric INAR(1) model," *Filomat*, 2017, doi: 10.2298/FIL1713009N.