FULafia Journal of Science & Technology
ISSN (print) 2449-0954 \ ISSN (online) 2636-4972
https://lafiascijournals.org.ng/index.php/fjst
Sponsored By tetfund TERTIARY EDUCATION TRUST FUND
Manuscript => Received: Dec., 2023; **Accepted:** March, 2024; **Published:** April, 2024 https://doi.org/10.62050/fjst2024.v8n1.281

EXPLORING THE INAR MODEL ON HEAVY TAILED TIME SERIES DATA WITH OUTLIERS

I. M. Saleh[1]*, I. Akeyede[1], A. M. Abubakar[2] and A. I. Sulaiman[2]

[1]*Department of Statistics, Federal University of Lafia, PMB 146 Lafia, Nigeria*
[2]*Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria*
*Corresponding email: saleh.musa@science.fulafia.edu.ng

ABSTRACT
Count data are intrinsically measures of event frequency; it is clear that there is an intrinsic relationship with recurring time to event. Events are typically tallied within time intervals for practical and convenient reasons. The existence of outliers is one issue that prevents count data from being stationary in time series analysis; this has an impact on the effectiveness of fitting several common stationary models to the count data collected over time. Thus, the purpose of this study was to examine how well the Integer Valued Autoregressive (INAR) model performed while modeling count data that contains outlier(s). While this model has been studied for count time series data, it has not been studied for varying degrees of outliers. A monte-carlo simulation was carried out to select the best INAR(p), where p=1, 2, 3 and 4 on data with 10, 20 and 30% outliers at different sample sizes. The INAR (4) has the best fit across the sample sizes at the larger percentages of outliers while INAR (3) at the lowest percentage with smallest information criteria of assessment and they are therefore recommended for such modeling.

Keywords: INAR(p), count data, outlier, modeling, simulation

INTRODUCTION
The values of certain statistical variables measured across a consistent collection of time points are called time series. Monthly sales in a store, monthly HIV/AIDS cases reported in a hospital, annual production by a corporation, the number of eggs laid daily by farm animals, the amount of power consumed in kilowatts, and information on daily population motor registration are a few examples of time series data. Counts, such the number of car accidents, hospital patients, customers waiting for service at a specific time, and so on, are frequently included in time series data.

In many fields, time series count data which is the number of times an item or event occurs within a specified period of time is crucial. These include the quantity of heart attacks or days spent in the hospital in medical studies, the quantity of absences from classes over a given period of time in education research, or the quantity of instances in which parents abuse their children in social science studies. In order to simulate the situation of a count random variable (RV) with a non-negative integer value, several statistical distributions have been utilized. An excellent summary of these distributions is provided by Johnson *et al.* (2005).

Time series of counts have been of interest to researchers, as evidenced by recent studies. For example, Weiß (2009) worked on time series of counts with overdispersion and suggests using Integer-Valued Generalized AutoregRessive Conditional Heteroskedasticity (INGARCH) models to describe the integer-valued processes with overdispersion. The work of the author was limited to the unique scenario where p = q = 1. In his analysis of several time series of claim counts for pay loss and health care-only claims at the Workers' Compensation Board of British Columbia (WCB), Harvey and Fernandes (1989) found that the time series of counts could be fitted by a stationary Poisson INAR(1) model. The indicated that the reported empirical mean and variance of the data, which are 8.60417 and 11.3575, respectively, are to blame for the genuine marginal distribution's overdispersion.

Furthermore, Freeland (1998) proposed that a Poisson INAR(1) model might not be the best option. This highlights the need for more model research. The work of Ndwiga *et al.* (2019) and Saleh *et al.* (2021) is another source of inspiration for this study. In it, he highlights the inappropriate application of the Traditional Generalized Linear Model (GLM) in the modeling of time series count data. He then examines the performance of various models, including Poisson, Negative Binomial, Zero-inflated Poisson, Hurdles Poisson, and Negative Binomial Hurdles. His findings suggest that Negative Binomial Hurdles outperformed other models in most scenarios, making it the most statistically fit model for overdispersed count data containing excess zeros.

In the quest for robust method for time series count data analysis (Akeyede *et al.*, 2022) considered modelling of heavy tailed count time series data on number of rotavirus data using heavy tailed probabilities, the researcher validate the capability of the model in accounting forover-dispersion though with some deficiency and further recommend the use of INAR or ACP model to analyse heavy tailed count time series data with outliers. Ndwiga *et al.* (2019) confirmed the in appropriate use of negative binomial distributions and poison distributions in modelling count time series especially with over dispersion, the researcher further

proposes the use of hurdle poison model for analysing data with over-dispersion or excess zeros. e.g., non-Gaussian clustered data, such as counts, are frequently modelled by making use of generalized linear mixed-effects models, which extend the broad class of generalized linear models by adding a subject specific random effect, often of a Gaussian type, to capture the correlation between the repeated measurements per subject, he notably pointed the need for a robust model capable of handling over-dispersion as well as under-dispersion as the case may be.

In practical applications, count data often exhibit outliers, over-dispersion, and even heavy-tailedness, where the tail probabilities are non-negligible or drop extremely slowly. While several models have been developed to model count data, Saleh *et al.* (2021), heavy-tailedness and presence of outliers have received less attention. Thus, the model an integer-valued autoregressive (INAR) process is intended to reflect this in this work. This study aimed at determining the accuracy of INAR model in modeling heavy tailed count time series data at different levels of outliers and sample sizes

MATERIALS AND METHODS

The INAR model is investigated on heavy tailed distributions and proportion of outliers. The effects of varying sample sizes (n=30,60,...,300) on the performance of the models were also examined. The optimal condition of orders p, with p = 1, 2, 3, and 4 correspondingly, is ascertained for the outlier levels at each sample size by the application of criteria such as the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Hannan-Quinn Information Criterion (HQIC). Data set were simulated in R statistical software with sample sizes of 30, 60, 90, …, 300, from poison and negative binomial distributions to produce count data with outliers. The model under study were then fitted to the simulated data so as to examine the effect of the proportion outliers on the selected orders of the INAR model. 0, 10% (low level outliers) and 20% (High level of outliers) of outliers were created on sample sizes of data simulated. Each constructed data set entails a specific effect of the outliers observed in the display of models outputs.

In simulation, we set our parameters to be $\emptyset_1 = 1 \emptyset_2 = 1$ to ensure discrete nature of count data generated. The response Y_{ti} in (1) were generated from poison and negative binomial distributions. The four model's orders under study were considered to analyze how well the model fits the selected data sets having some proportions of outliers.

Data were generated from linear second orders of autoregressive functions given as follows:

Model 1. AR(2): $Y_{ti} = Y_{ti-1} + Y_{ti-2} + e_t$ (1)

$t = 30, 60, 90, 120, 150, 180, 210, 240, 270, 300.$

$i = 1,2, \dots, 1000$

Where Y_{ti} were simulated from poison families for outliers as follows:

$$\frac{\lambda^{y_i} e^{-\lambda_i}}{y_i!}, for y_i = 0,1,2, \dots \quad (2)$$

Thus, for the Poisson models $E(y_i) = V(y_i) = \mu_i$.

Also, Y_{ti} were simulated from negative binomial families for outliers as follow

$$p(y_i; \lambda_i, \alpha_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha_i}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha_i}\right)} \left(\frac{1}{1 + \alpha_i \lambda_i}\right)^{\frac{1}{\alpha_i}} \left(\frac{\alpha_i \lambda_i}{1 + \alpha_i \lambda_i}\right)^{y_i}, i = 1, \dots, 1000 \quad (3)$$

Here, the dispersion parameter $\alpha_i > 0, \lambda_i = E(Y_i)$; and $V(Y_i) = \lambda_i + \alpha_i \lambda_i^2$

Integer-valued autoregressive (INAR) model

In this work we are interested in a special class models, the so-called integer-valued autoregressive (INAR) process introduced by McKenzie (1985), Al-Osh and Alzaid (1987). The theoretical properties and practical applications of INAR and related processes have been discussed extensively in the literature. Silva *et al.* (2005) consider independent replications of count time series modelled by INAR and proposed several estimation methods using the classical and Bayesian approaches in time and frequency domains.

Point prediction for INAR process

Suppose a non-negative integer-valued random variable X and λ ∈ [0, 1], the generalized relation which is denoted by $\lambda\ (X)$, is given by;

$$\lambda\ (X) = \sum_{j=1}^{X} Y_j \quad (4)$$

where {Y_j}, j =1, ..., X, is a sequence of independent and identically distributed non-negative integer-valued random variables, independent of X, with finite mean λ and variance σ^2. The sequence is known as the counting series of $\lambda\ (X)$. When {Yj} is a sequence of Bernoulli random variables, the thinning operation is called binomial thinning operation and was defined by Steutel and van Harn (1979). The well-known INAR(1) process $\{X_t; t = 0, \pm 1, \pm 2, \dots\}$ is defined on the discrete support $\mathbb{N}_o$ by the equation

$$Xt = \lambda X_{t-1} + \varepsilon_t \quad (5)$$

where $0 < \lambda < 1, \{\varepsilon_t\}$ is a sequence of independent and identically distributed integer-valued random variables, with $E[\varepsilon_t] = \mu_\varepsilon and Var[\varepsilon_t] = {\sigma_\varepsilon}^2$.Suppose a non-negative integer-valued random variable X and λ ∈ [0, 1], the generalized thinning operation which is denoted by '∘', is given by;

$$\lambda \circ X = \sum_{j=1}^{X} Y_j \quad (6)$$

where {Yj}, j =1, ..., X, is a sequence of independent and identically distributed non-negative integer-valued random variables, independent of X, with finite mean λ and variance σ^2. The sequence is known as the counting series of $\lambda \circ$ X. When {Y_j} is a sequence of Bernoulli random variables, the thinning operation is

called binomial thinning operation and was defined by Steutel and van Harn (1979). The well-known INAR(1) process $\{X_t;\ t = 0, \pm1, \pm2, \dots\}$ is defined on the discrete support $\mathbb{N}_o$ by the equation

$$Xt = \lambda \circ X_{t-1} + \varepsilon_t \qquad (7)$$

Where $0 < \lambda < 1, \{\varepsilon_t\}$ is a sequence of independent and identically distributed integer-valued random variables, with $E[\varepsilon_t] = \mu_\varepsilon and Var[\varepsilon_t] = {\sigma_\varepsilon}^2$.

Four orders: INAR (1), INAR (2), INAR (3) and INAR (4) models were considered in this work. To achieve this, R-code were developed for their estimation in R. The results of the analyses were be compared and presented in tables and graphs. The model with the least criteria of AIC, BIC and HQIC are considered to be the best for the data with various outliers and sample sizes.

RESULTS AND DISCUSSION

The performance of INAR models were determined through simulations on the count data that contain outliers. The effect of sample sizes $n = 30, 100, \dots, 300$, on the performance of the models were studied. At every sample size, the best status of the p, where p = 1, 2, 3, 4 are respectively determined for the levels of the outliers in the data generated using criteria like AIC, BIC and HQIC as presented in Tables 1 – 3 and plotted on Figures 1a – 3c. 0, 10 and 20% of outliers representing no, low and high levels of outliers respectively were injected in the data so as to determine the best INAR model for each category. The simulation study was carried out with 1000 iteration on each case in R statistical software. For each iteration, the values of the criteria for the assessment (AIC, BIC and HQIC) were computed and their average values were recorded according to sample sizes as shown in Tables 1 – 3. The values from the tables were plot in Figures 1a–3c. The model with lowest criteria is considered as the best.

Table 1: Performance of fitted models with no outlier (0%)

Criteria	Sample Sizes/ Model	30	60	90	120	150	180	210	240	270	300
AIC	INAR(1)	-53.21	-122.42	-135.03	-220.49	-243.93	-328.15	-362.69	-435.76	-509.04	-612.13
	INAR(2)	-56.59	-127.49	-159.31	-228.3	-264.21	-350.7	-394.98	-468.65	-540.14	-682.88
	INAR(3)	-42.79	-122.76	-149.05	-222.33	-249.5	-321.66	-383.2	-429.65	-524.76	-606.25
	INAR(4)	-59.22	-134.03	-169.8	-250.51	-271.44	-365.12	-399.49	-444.37	-556.58	-618.10
BIC	INAR (1)	55.326	124.485	187.083	252.527	265.967	360.174	454.71	437.781	601.056	674.152
	INAR (2)	60.818	115.629	163.410	222.381	258.273	334.759	359.03	432.694	544.182	586.913
	INAR (3)	49.133	128.961	155.203	228.447	265.603	337.747	389.27	435.719	530.817	632.307
	INAR (4)	53.669	112.307	138.010	188.671	249.575	303.237	350.59	422.462	524.660	626.175
HQIC	INAR(1)	58.255	113.682	158.22	198.201	229.432	279.271	372.84	342.131	466.901	486.624
	INAR(2)	46.630	105.417	146.663	167.322	227.383	259.807	329.00	344.881	407.259	454.401
	INAR(3)	41.708	102.347	144.802	173.535	231.256	281.795	336.34	351.731	399.786	436.413
	INAR(4)	34.976	95.882	123.454	159.258	221.786	238.939	315.44	325.961	390.732	414.011

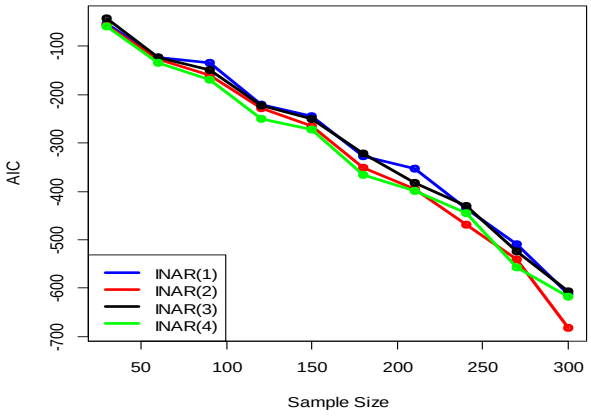

Figure 1a: AIC of the fitted INAR (p) models when there is no outlier

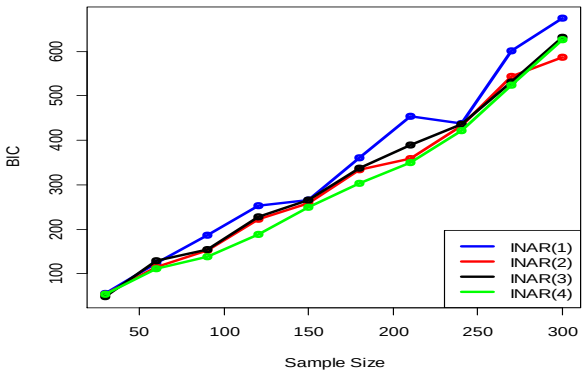

Figure 1b: BIC of the fitted INAR (p) models when there is no outlier

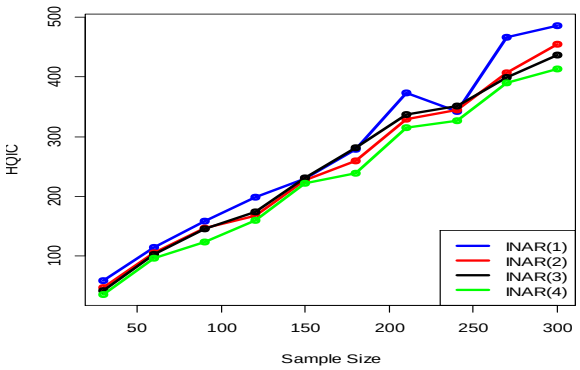

Figure 1c: HQIC of the fitted INAR (p) models when there is no outlier

Table 1 shows the average values of AIC, BIC and HQIC of the fitted models computed from 1000 iteration of data simulated from poison distribution without outlier. The values of the criteria in the Table 1 were plotted on Figures 1a, 1b and 1c, respectively. It is observed that the model's orders exhibit similar pattern of fit across sample sizes. However the best fitted model is INAR (4) followed by INAR (2) with least value three criteria across sample sizes. INAR (1) seemingly show the least fit for count time series data with no outlier.

Table 2: Performance of fitted models with low level of outliers (10%)

Criteria	Sample Sizes/ Model	30	60	90	120	150	180	210	240	270	300
AIC	INAR(1)	-63.97	-121.23	-192.18	-253.00	-299.31	-360.71	-440.33	-539.01	-549.50	-618.33
	INAR(2)	-57.87	-118.66	-171.34	-250.81	-298.97	-344.92	-416.11	-501.80	-535.90	-610.98
	INAR(3)	-52.89	-116.36	-165.05	-245.70	-288.15	-344.87	-397.57	-499.92	-513.40	-598.46
	INAR(4)	-49.36	-111.31	-157.21	-243.96	-283.79	-336.69	-389.41	-498.57	-510.90	-581.85
BIC	INAR (1)	66.083	123.302	194.234	255.044	301.346	362.735	442.35	541.036	551.554	620.348
	INAR (2)	62.091	122.796	175.435	254.886	303.039	348.973	420.16	505.846	539.973	615.020
	INAR (3)	59.226	122.561	171.2	251.824	294.245	350.956	403.64	505.988	519.464	604.512
	INAR (4)	57.816	119.581	165.406	252.115	291.921	344.803	397.49	506.659	519.006	589.929
HQIC	INAR(1)	65.663	130.887	202.165	251.476	310.520	412.470	463.92	553.919	565.459	641.312
	INAR(2)	52.462	121.306	177.308	235.33	300.420	379.290	416.06	510.627	519.296	601.269
	INAR(3)	51.469	121.078	174.752	232.652	298.050	377.800	413.01	505.237	515.528	599.851
	INAR(4)	50.724	119.86	173.362	230.011	295.760	376.570	411.89	505.162	515.398	594.688

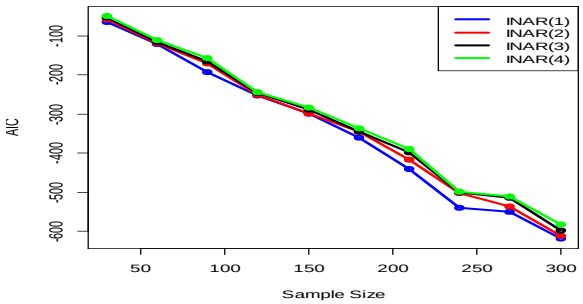

Figure 2a: AIC of the fitted INAR (p) models when there is low proportion outlier

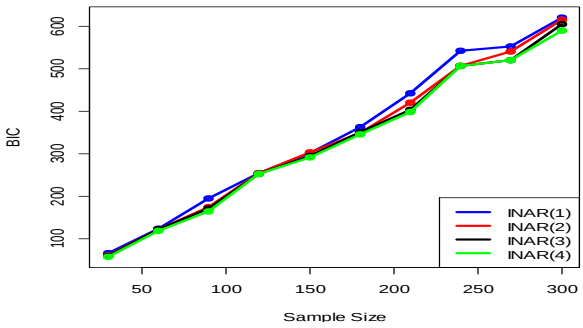

Figure 2b: BIC of the fitted INAR (p) models when there is low proportion of outliers

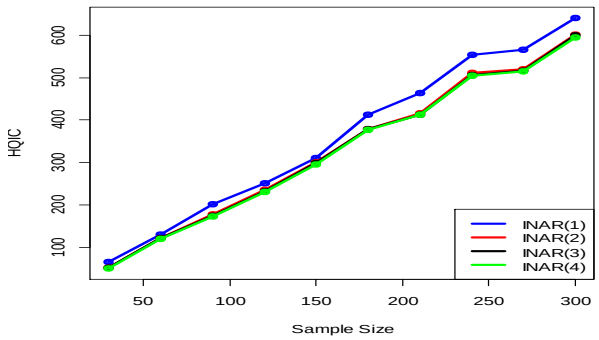

Figure 2c: HQIC of the fitted INAR (p) Models when there is low proportion of outliers

The average values of the fitted models' AIC, BIC, and HQIC which were computed from 1000 iterations of data simulated from a poison distribution without 10% outlier are displayed in Table 2. Plots of the criteria values from Table 2 were made in Figures 2a, 2b, and 2c, in that order. It is noted that the model's orders show a consistent fit pattern for all sample sizes. With the least value three criteria across sample sizes, INAR (2) is the second best fitted model, after INAR (4). The least fit for count time series data without low outlier appears to be indicated by INAR (1).

Table 3: Performance of fitted models with high level of outliers (20%)

Criteria	Sample Sizes/ Model	30	60	90	120	150	180	210	240	270	300
AIC	INAR(1)	-69.92	-125.59	-177.96	-254.08	-291.62	-357.97	-454.00	-526.15	-546.72	-622.23
	INAR(2)	-54.39	-121.89	-176.63	-247.62	-284.63	-337.10	-413.46	-496.89	-545.52	-617.23
	INAR(3)	-19.23	-87.82	-151.41	-201.45	-286.71	-322.19	-392.83	-482.69	-537.74	-607.73
	INAR(4)	-27.00	-95.631	-165.36	-223.29	-249.23	-306.36	-390.17	-478.85	-531.67	-606.28
BIC	INAR (1)	72.031	127.656	180.01	256.115	293.649	359.993	456.02	528.174	548.743	624.247
	INAR (2)	58.612	126.021	180.73	251.703	288.691	341.159	417.51	500.936	549.563	621.271
	INAR (3)	25.574	101.021	170.56	207.564	292.811	328.273	398.90	488.767	543.798	613.782
	INAR (4)	35.453	103.904	173.56	231.454	257.362	314.472	398.27	486.945	539.749	614.358
HQIC	INAR(1)	53.711	120.866	162.265	244.577	307.65	377.75	412.50	496.384	526.866	524.999
	INAR(2)	60.889	122.413	178.581	228.411	279.95	324.68	366.66	450.354	488.872	497.698
	INAR(3)	24.556	111.554	166.720	198.811	278.62	322.61	360.46	448.281	487.738	496.568
	INAR(4)	49.567	113.717	168.517	218.534	267.21	320.76	359.97	447.568	486.846	496.045

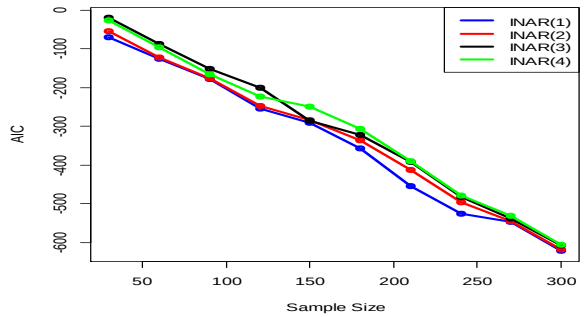

Figure 3a: AIC of the fitted INAR (p) models when there is high proportion of outliers

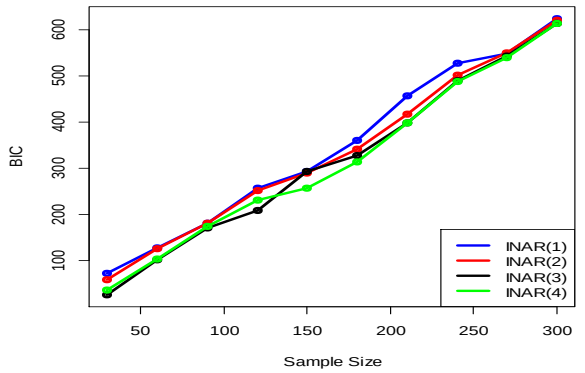

Figure 3b: BIC of the fitted INAR (p) models when there is high proportion of outliers

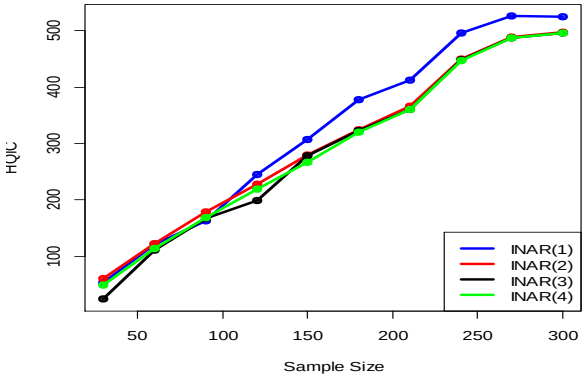

Figure 3c: HQIC of the fitted INAR (p) models when there is high proportion of outliers

Plots of the criteria values from Table 2 were made in Figures 2a, 2b, and 2c, respectively. INAR (3) is more robust to higher level of outlier compare to other models especially when sample size is low and moderate 200, as sample size increase, the INAR (4) take the lead with the least value BIC and HQIC criteria. The least fit for count time series data high level of outlier appears to be indicated by INAR (3) and INAR (4) at small and large sample size, respectively.

CONCLUSION

A problem that is frequently encountered in many scientific and public health applications is time series of count with outlier. This kind of series has proven to be quite challenging to statistically model. As a result, our investigation has identified a model with a few outlier levels. When evaluating such data, failing to take these into account might lead to the identification of spurious connections as well as inaccurate and occasionally misleading conclusions. INAR (1), INAR (2), INAR (3) and INAR (4) were used to fit data with no outlier, 10% outliers and 20% outliers which represent no, low and high level of outliers in order to determine the best among the aforementioned models for each category of outliers injected in the simulated data and sample size. It was discovered that the highest performing model when fitted different count time series data with different levels of outlier was the INAR (4) when there is no outlier in the data and when there is low level of outliers at various sample sizes. However, INAR (3) is more robust to higher level of outlier compare to other models especially when sample size is low and moderate 200, as sample size increase, the INAR (4) take the lead with the least value BIC and HQIC criteria. The least fit for count time series data high level of outlier appears to be indicated by INAR (3) and INAR (4) at small and large sample size respectively. This work suggested some specific time series models that can be used to fit a type of count data with some accompanying outlier structure.

REFERENCES

Akeyede, I., Bakari, H. R. and Muhammad, R. B. (2022). Robustness of ARIMA and ACP models to over-dispersion in analysis of count data. *Journal of Nigeria Statistical Association*, 34, 95-105. https://nsang.org/_uploads/uploads/2021/60112b447fdc2_5.pdf

Freeland R. K. (1998). Statistical analysis of discrete time series with applications to the analysis of workers compensation claims data. PhD Thesis, University of British Columbia, Canada. Retrieved from https://dx.doi.org/10.14288/1.0088709

Harvey, A. & Fernandes, C. (1989). Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*, 7(4), 407-417. doi:10.2307/1391639

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distribution*. Wiley, New York, DOI:10.1002/0471715816

Musa, S. I., Nweze, N. O. and Adenomon, M. O. (2021). On performance of integer-valued autoregressive and poisson autoregressive models in fitting and forecasting time series count data with excess zeros. AJMS, 5(2), 20-27. https://doi.org/10.22377/ajms.v5i2.322

Ndwiga A. Macharia, Oscar Ngesa, Anthony Wanjoya and Damaris FelistusMulwa (2019). Comparison of Statistical Models in ModelingOverDispersed Count Data with Excess Zeros. *International Journal of Research and Innovation in Applied Science* (IJRIAS), IV(V), 80-90. URL:https://www.rsisinternational.org/journals/ijrias/DigitalLibrary/Vol.4&Issue5/80-90.pdf

Popovic, P. M. (2015). A bivariate INAR(1) model with different thinning parameters. *Statistical Papers*, DOI 10.1007/s00362-015-0667-1.

Saleh, I. M. and N. O. Nweze (2021). Model selection for time series count data with over dispersion. *Asian Journal of Probability and Statistics*, 14(2), 60-73. https://doi.org/10.9734/ajpas/2021/v14i230326

Silva, I., Silva, M. E., Pereira, I. and Silva, N. (2005). Replicated INAR(1) process. *Methodology and Computing in Applied Probability*, 7, 517–542. DOI:10.1007/s11009-005-5006-x

Steutel, F. W. and Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 5, 893–899. DOI: 10.1214/aop/1176994950

Weiß, H. (2009). Modelling time series of counts with over dispersion. *Stat Methods Appl.*, 18, 507–519. DOI 10.1007/s10260-008-0108-6