Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/150403

Modelling epidemiological count time series using Integer-valued GARCH models embedded with Fourier terms

Thembhani Hlayisani Chavalala[1,2,*] , *Retius Chifurira*[2], *Knowledge Chinhamu*[2], *and Jacob Majakwara*[3]

[1]Department of Statistics and Operations Research, University of Limpopo, South Africa
[2]School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, South Africa
[3]School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

Received: 3 Apr. 2026, Revised: 15 May 2026, Accepted: 27 Jun. 2026
Published online: 1 Jul. 2026

Abstract: Frequently, epidemiological count data are characterised by complex features such as overdispersion, serial autocorrelation, heteroscedasticity, and seasonal patterns. Poisson and Negative Binomial (NB)-INGARCH models have been extensively utilised to study and forecast datasets of this nature. These modelling frameworks do not account for the seasonal patterns that are often exhibited by the epidemiological time series. The inability to capture seasonality in the count time series process could lead to biased forecasts, misinterpretation of the underlying trends, and inadequate model performance. This, in turn, undermines the efficient allocation of essential resources such as healthcare personnel and vaccination programs. To ensure sound healthcare planning and timely interventions, effective modelling of epidemiological data remains fundamental. This study incorporates the Fourier components in the NB-INGARCH model to address the seasonality often exhibited in epidemiological data. For comparative analysis, various INGARCH models were fitted to the daily number of COVID-19 death cases in South Africa, and as well as in its three largest Provinces. The NB-INGARCH framework consistently outperformed the Poisson-INGARCH models, with the inclusion of Fourier terms further enhancing its effectiveness. Specifically, the NB-INGARCH(2,1) model incorporating Fourier harmonics successfully captured overdispersion, short-term dependence, heteroscedasticity, and weekly seasonality. The findings highlight that uniform seasonal adjustments may be inadequate, underscoring the need for tailored, region-specific modifications. A one-size-fits-all modelling approach risks oversimplifying the complexity and variability inherent in epidemiological count time series.

Keywords: Epidemiology, COVID-19, Poisson-INGARCH, NB-INGARCH, Fourier harmonic terms

1 Introduction

Epidemiological count time series present unique statistical challenges because they consist of non-negative integer observations that frequently exhibit serial dependence, overdispersion, excess variability, and abrupt fluctuations associated with epidemic waves. During large-scale infectious disease outbreaks such as the COVID-19 pandemic, these characteristics become particularly pronounced, resulting in periods of persistent increases and decreases in disease incidence or mortality that resemble the volatility clustering commonly observed in financial return series. In financial econometrics, volatility clustering, whereby large changes tend to be followed by large changes and small changes by small changes, has been extensively modelled using autoregressive conditional heteroscedasticity (ARCH) and generalized autoregressive conditional heteroscedasticity (GARCH) frameworks and their numerous extensions [1–5]. Although epidemiological data are fundamentally different from financial returns because they are discrete counts rather than continuous-valued observations, epidemic processes often generate analogous conditional variance dynamics arising from successive waves of disease transmission, public health interventions, behavioural changes, and reporting delays. Consequently, daily COVID-19 death counts in South Africa, both nationally and within its three largest provinces (Gauteng, KwaZulu-Natal and Western Cape), exhibit overdispersion, serial correlation and autoregressive

* Corresponding author e-mail: thembhani.chavalala@ul.ac.za

conditional heteroscedasticity (ARCH) effects indicative of volatility clustering [6].

These empirical characteristics necessitate the use of modelling frameworks capable of jointly capturing the discrete nature of the data, temporal dependence and time-varying conditional variability. The integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) model provides a natural extension of the ARCH/GARCH paradigm to count time series by modelling the conditional mean as a function of past observations and past conditional means while accommodating overdispersion and serial dependence [7]. As such, INGARCH models offer a theoretically sound and practically relevant framework for analysing epidemic count processes exhibiting volatility clustering analogous to that observed in financial time series.

The INGARCH modelling framework serves as the integer-valued analogue of the generalised ARCH (GARCH) model which has been highly useful in capturing the volatility dynamics within finance and economics [8]. In the recent decades, the INGARCH model has garnered considerable attention, driven by the increasing need to model time series of counts across diverse domain such as finance and economics (e.g., daily transaction volumes at a bank) and epidemiology (e.g., daily hospital admissions for a particular disease) [9, 10]. The Poisson-based INGARCH model introduced by [11], is the most utilised count time series modelling framework [9, 12, 13]. However, Poisson distribution is well-suited for equi-dispersed dataset, where the mean and variance of the process are equal [14–17]. The equi-dispersion property often poses challenges in modelling count time series because the integer-valued data frequently exhibit overdispersion [14, 15, 18].

There are numerous alternative discrete probability distributions that can be utilised to model overdispersed time series processes, where the mean is less than the variance of the time series [16]. Zhu [19] introduced the NB-INGARCH model as a time series framework for integer-valued data. The NB distribution extends the Poisson framework by incorporating an additional dispersion parameter [17]. The dispersion parameter enhances the flexibility and robustness in modelling count processes with variability exceeding that of the Poisson framework. In essence, the NB-INGARCH framework is designed to jointly handle overdispersion while mitigating the influence of potential outliers in count data [16, 19]. However, empirical data in epidemiology exhibit other complex dynamics such as strong seasonal and cyclical patterns that cannot be adequately explained by the overdispersion parameter and autoregressive terms alone [20]. Accordingly, it becomes necessary to extend the NB-INGARCH framework to incorporate the Fourier harmonic terms that can effectively account for seasonality, overdispersion, temporal dependence, and the presence of potential influential outliers.

Fourier harmonic terms are constructed from sine and cosine functions within the Fourier series to provide a parsimonious way of modelling seasonality in the data [21]. According to [22], Fourier series effectively addresses the complexity of seasonal patterns in the data while producing superior predictive accuracy. Letema and Mbwambo [23] incorporated the Fourier terms to account for seasonality in their study of modelling and forecasting the tourist arrivals in Tanzania. Therefore, integrating the Fourier harmonic components within the NB-INGARCH framework enhances the model to simultaneously account for temporal dependence, overdispersion, seasonality and extreme observations. This inference is based on the suggestion by [22], who argued that the inclusion of Fourier terms boosts forecasting accuracy by uncovering latent cyclical structures that might otherwise remain undetected. Meanwhile, the NB-INGARCH component ensures appropriate modelling of the conditional variance structure.

Numerous modelling techniques for inter-valued data have been proposed and studied [10, 24, 25]. Nevertheless, to the best of our knowledge, the application of an integer-valued GARCH framework augmented with Fourier components to capture seasonality has not yet been explored in epidemiological contexts, especially in South Africa. Therefore, this study seeks to find a well-suited INGARCH modelling framework that integrates Fourier harmonic terms for modelling and forecasting seasonal epidemiological count time series in South Africa.

2 Data and Methodology

2.1 Data

The study will utilise time series data of COVID-19 death cases in South Africa, as well as in its three populous provinces to explore the integration of Fourier harmonic components within the INGARCH modelling framework. The three provinces were considered the primary hotspots of the COVID-19 pandemic in South Africa. Daily data were collected and published by the National Institute for Communicable Diseases (NICD) in South Africa via their official website, and are freely accessible at: https://github.com/dsfsi/covid19za/blob/master/data/. Figure 1 depicts the time series plots for the daily death counts of COVID-19 collected over the period from 27 March 2020 to 22 July 2022.

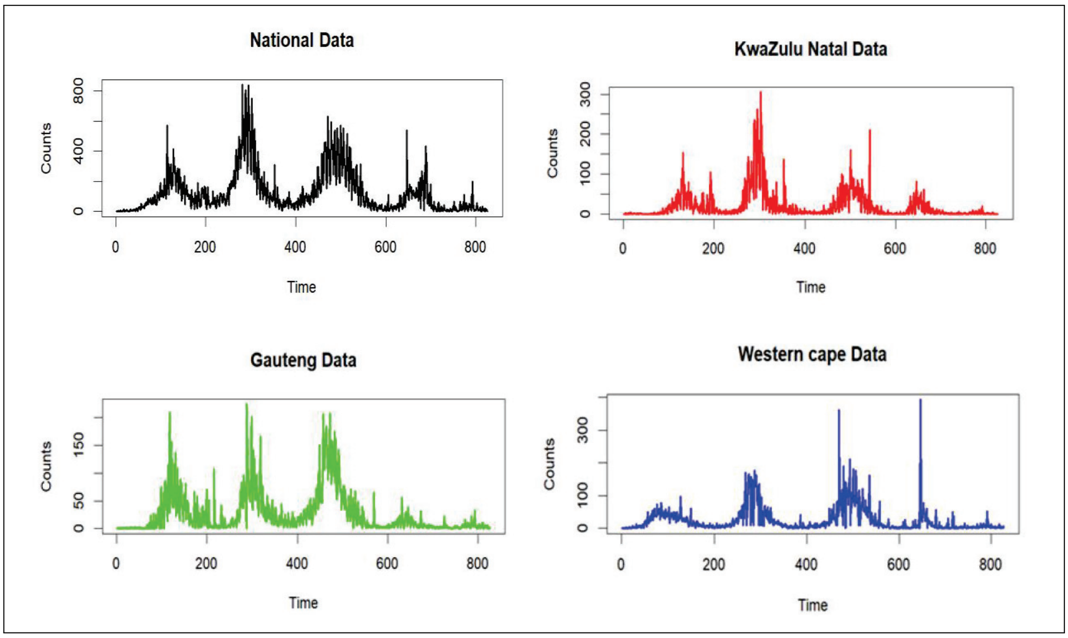

Fig. 1: Daily COVID-19 mortality cases in South Africa, and in its three epicenters

The four plots in Figure 1 depicts that all the considered time series data exhibit similar patterns, with periods of low and high counts occurring concurrently. These temporal fluctuations could be linked to the emergence of different waves (variants). The presence of zeros in the data leads to mean value that is lower than the corresponding variance. To validate the characteristics of the COVID-19 mortality cases that are virtually apparent in the plots, additional exploratory data analysis (descriptive statistics and hypothesis testing) was conducted, with the findings presented in Table 1.

Panel A in Table 1 shows that the mean of counts is consistently lower than the corresponding variance across all the four datasets, resulting in dispersion indices (variance-to-mean ratios) significantly above one. These values support the inference deduced from Figure 1 that the time series of COVID-19 death counts are overdispersed. In Panel B, the three tests for serial autocorrelation reveals that the daily deaths due to COVID-19 are highly autocorrelated because the p-values of the test statistics are significantly smaller than the 5% level. This implies that the past observation in the count time series contain useful information for forecasting the future value, thereby confirming the presence of temporal dependence in the data. The Lagrange Multiplier test statistic for ARCH effects in Panel C have corresponding p-values substantially lower than the 5% level of significance. These small p-values lead to the rejection of the null hypothesis which say there is no ARCH effects in the data, thereby verifying the presence of ARCH effects in the data. That is, volatility varies over time in the daily death counts of COVID-19 in South Africa and its three largest Provinces.

The daily COVID-19 deaths cases, as depicted in Figure 1 and substantiated by descriptive statistics and tests in Table 1, may be characterised as overdispersed, highly autocorrelated, non-negative integer values exhibiting volatility clustering. Discrete count time series frameworks have been developed to accurately model datasets with characteristics like the ones observed in the South African COVID-19 data [17,26,27]. Mubanga [17] conducted a comparative analysis study of the Poisson and NB-INGARCH models within the context of malaria epidemiology in Zambia. However, there is a need to integrate the Fourier harmonic components in these INGARCH models so that they can effectively account for seasonality. Failure to account for the seasonal cycles in epidemiology modelling could lead to inaccurate forecasts, thereby compromising the effective allocation of critical resources, including hospital staffing and vaccination programs. Thus, this study aims to close this gap by incorporating the Fourier terms in the INGARCH framework.

Table 1: Exploratory data analysis of the COVID-19 mortality cases in South Africa, and in its three epicenters

Panel A: Summary statistics of the datasets

	National	Gauteng	KwaZulu-Natal	Western Cape
Mean	123.2684	25.4051	19.6675	26.9637
Variance	20094.97	1415.304	1354.564	1660.919
Proportion of zeros	0.01451	0.13422	0.20677	0.11971
Dispersion index	163.0180	55.6701	68.8733	55.4310
Minimum	0	0	0	0
Maximum	844	306	225	394

Panel B: Testing the null hypothesis: No Autocorrelation

	National		KwaZulu-Natal		Gauteng		Western Cape	
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Durbin-Watson	0.3623	< 0.0001	0.5765	< 0.0001	0.5359	< 0.0001	0.8700	< 0.0001
Box-Pierce	4546.9	< 0.0001	2917.9	< 0.0001	3782.1	< 0.0001	2399.4	< 0.0001
Box-Ljung	4587.8	< 0.0001	2943.5	< 0.0001	3816.0	< 0.0001	2420.8	< 0.0001

Panel C: Testing the null hypothesis: No ARCH Effects

	National		KwaZulu-Natal		Gauteng		Western Cape	
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Lagrange Multiplier	580.7	< 0.0001	481.61	< 0.0001	375.64	< 0.0001	109.6	< 0.0001

2.2 Methodology

Figure 2 illustrates the flowchart of the modelling framework developed in this study.

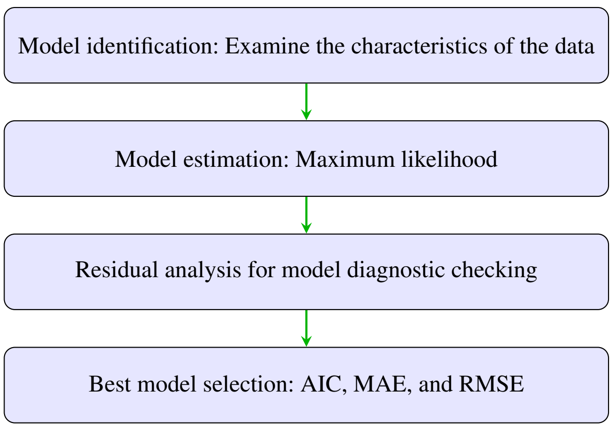

Fig. 2: Step by step workflow diagram of the proposed modelling framework

2.2.1 The INGARCH model

The INGARCH model is considered as the integer-valued analogue of the traditional generalised autoregressive conditional heteroscedastic (GARCH) framework introduced by [26]. The GARCH framework was originally designed to model the dependence of the current variances on their past values, whereas the INGARCH model extends this idea to the discrete domain by accurately forecasting the conditional mean values using the past mean values and previous observations.

2.2.1.1 Poisson-INGARCH model .

Suppose that $\{Y_t\}_{t\in\mathbb{Z}}$ is a non-negative integer time series process, and let $p \geq 1$ and $q \geq 0$. The time series process Y_t follows a Poisson-INGARCH(p,q) model if it satisfies the following conditions:

$$\begin{cases} (i)\ Y_t \mid \mathscr{F}_{t-1} \sim \text{Poisson}(\lambda_t) \quad \forall t \in \mathbb{Z} \\ (ii)\ \lambda_t = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \end{cases} \tag{1}$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, for $1 \leq i \leq p$, $1 \leq j \leq q$, and $\mathscr{F}_{t-1}$ is the σ-field generated by $\{Y_k : k \leq t\}$. The INGARCH(p,q) process is said to be stationary if $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$.

The most popular shortcoming of model (1) is the requirement that the conditional mean must always be equal to the conditional variance. This shortcoming is frequently encountered in epidemiological data, and it could lead to poor performance especially in the presence of potential extreme observations [19]. There are several alternative distributions that can be employed to model datasets with variance greater than the mean [16]. The NB-INGARCH model was introduced as the generalisation of model (1) to accommodate more overdispersed data. However, this model does not address the issues of serial autocorrelation that sometimes exists in the time series data.

2.2.1.2 NB-INGARCH model .

Suppose that the discrete count time series Y_t follows a NB distribution with parameters r and p_t. Formally, it can be expressed as

$$Y_t \mid \mathscr{F}_{t-1} \sim \text{NB}(r, p_t) \quad \forall t \in \mathbb{Z}, \tag{2}$$

where $\mathscr{F}_{t-1}$ is the σ-field generated by $\{Y_k : k \leq t\}$, r is a non-negative number, and p_t satisfies the regression equation.

$$\lambda_t = \frac{1-p_t}{p_t} = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \tag{3}$$

with constraints $\omega > 0$, $\alpha_i \geq 0$, and $\beta_j \geq 0$, for $1 \leq i \leq p$ and $1 \leq j \leq q$. The specifications in (3) defines the NB-INGARCH(p,q) process. The conditional probability mass function of Y_t is given by

$$P(Y_t = y_t \mid \mathscr{F}_{t-1}) = \binom{y_t + r - 1}{r-1} p_t^r (1-p_t)^{y_t} \quad y_t = 0, 1, 2, \ldots. \tag{4}$$

Since $p_t = \frac{1}{1+\lambda_t}$, and $1 - p_t = \frac{\lambda_t}{1+\lambda_t}$, the probability mass function can be equivalently expressed as

$$P(Y_t = y_t \mid \mathscr{F}_{t-1}) = \binom{y_t + r - 1}{r-1} \left(\frac{1}{1+\lambda_t}\right)^r \left(\frac{\lambda_t}{1+\lambda_t}\right)^{y_t} \quad y_t = 0, 1, 2, \ldots. \tag{5}$$

The conditional mean and variance of the count time series Y_t are given by

$$\begin{cases} E(Y_t \mid \mathscr{F}_{t-1}) = \frac{r(1-p_t)}{p_t} = r\lambda_t \\ \text{Var}(Y_t \mid \mathscr{F}_{t-1}) = \frac{r(1-p_t)}{p_t^2} = r\lambda_t(1+\lambda_t). \end{cases} \tag{6}$$

Clearly, the variance exceeds the mean, i.e., $\text{Var}(Y_t \mid \mathscr{F}_{t-1}) > E(Y_t \mid \mathscr{F}_{t-1})$. This inequality can also be demonstrated through the variance decomposition:

$$\text{Var}(Y_t) = E[\text{Var}(Y_t \mid \mathscr{F}_{t-1})] + \text{Var}(E(Y_t \mid \mathscr{F}_{t-1})).$$

Substituting the expressions in (6) yields

$$\text{Var}(Y_t) = E[r\lambda_t(1+\lambda_t)] + \text{Var}(r\lambda_t) = rE(\lambda_t) + r\big(E(\lambda_t)\big)^2 + (r+r^2)\,\text{Var}(\lambda_t),$$

which is strictly greater than

$$rE(\lambda_t) = E(\lambda_t). \tag{7}$$

Highlighting the overdispersion property inherent in the NB-INGARCH process.

2.2.1.3 INGARCH model with Fourier terms .

The NB-INGARCH model was introduced to capture both the overdispersion and potential extreme observations [19]. However, there are situations where the count data is overdispersed with some extreme observations and seasonal fluctuations. In order to deal with the seasonal fluctuations in the data, the NB-INGARCH can be extended by incorporating the Fourier terms as the deterministic covariates in the conditional mean structure.

Suppose that Y_t denote a non-negative integer time series process for all $t \in \mathbb{Z}$. The NB-INGARCH(p,q) model with K seasonal harmonics of period S specifies the conditional mean $\lambda_t = E(Y_t \mid \mathscr{F}_{t-1}) = \frac{1-p_t}{p_t}$ as a linear function of the p past observations and q previous conditional means:

$$\lambda_t = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} + \sum_{k=1}^{K} \left\{ \gamma_k \sin\left(\tfrac{2\pi kt}{S}\right) + \delta_k \cos\left(\tfrac{2\pi kt}{S}\right) \right\} \quad t \in \mathbb{Z}, \tag{8}$$

where $\omega > 0$ is the baseline parameter, $\alpha_i \geq 0, \quad i = 1,2,\ldots,p$ are the autoregressive coefficients, $\beta_j \geq 0, \quad j = 1,2,\ldots,q$ are the moving-average coefficients, $\gamma_k \in \mathbb{R}$, and $\delta_k \in \mathbb{R}$ for $k = 1,2,\ldots,K$ are the regression coefficients of the k^{th} sine and cosine Fourier harmonic terms, respectively. The seasonal component of the model can be expressed as

$$f(t) = \sum_{k=1}^{K} \left\{ \gamma_k \sin\left(\tfrac{2\pi kt}{S}\right) + \delta_k \cos\left(\tfrac{2\pi kt}{S}\right) \right\} \quad t \in \mathbb{Z},$$

where the number of sine and cosine harmonic terms K is determined by the seasonal period S. That is, the maximum number of harmonic terms that can be included in the model is computed by $K = S/2$. Even though, the formula gives the maximum, the appropriate number of harmonic terms is determined based on the lower Akaike information criterion (AIC) value [29].

The necessary and sufficient stationarity condition in the mean of the NB-INGARCH(p,q) process with Fourier terms is that all the roots $z_i, \quad i = 1,2,\ldots,p$ of the polynomial equation

$$1 - \sum_{i=1}^{q} (r\alpha_i + \beta_i) z^{-i} - \sum_{i=q+1}^{p} r\alpha_i z^{-i} = 0, \tag{9}$$

lie inside the unit circle. The harmonic terms are deterministic regressors representing the seasonal cycles which allow the NB-INGARCH process to account simultaneously for the autocorrelation, overdispersion, and seasonality exhibited by the discrete time series data. They do not change the stochastic stationarity condition since they are bounded function of time:

$$\left| \gamma_k \sin\left(\tfrac{2\pi kt}{S}\right) + \delta_k \cos\left(\tfrac{2\pi kt}{S}\right) \right| \;\leq\; \sqrt{\gamma_k^2 + \delta_k^2}. \tag{10}$$

The integration of the Fourier terms ensures that the recurrent peaks and troughs in the data are modelled explicitly, reducing the residual autocorrelation at seasonal lags and improving both the out-of-sample and in-sample forecast accuracy.

2.2.1.4 ZINBA-GARCH model .

For comparative analysis, we adopt a hybrid modelling framework that integrates the Zero Inflated Negative Binomial Autoregressive (ZINBA) specification with a GARCH(1,1) process. The ZINBA component effectively addresses overdispersion and the presence of excess zeros, depicted in Figure 1 and Table 1. In turn, the GARCH(1,1) structure captures time varying volatility and ARCH effects within the residuals of the ZINBA model. By embedding the residual process in a GARCH(1,1) framework, the proposed ZINBA GARCH(1,1) model provides a comprehensive representation that simultaneously accounts for the distributional dynamics of count data and the conditional heteroskedasticity inherent in its residuals [6]. The volatility equation is expressed as

$$\sigma_t^2 = \omega + \alpha \eta_{t-1}^2 + \beta \sigma_{t-1}^2, \quad \eta_t = \sigma_t z_t, \quad z_t \sim N(0,1), \tag{11}$$

where η_{t-1} denotes the standardized residuals of the ZINBA model, $\omega > 0$ is the constant term, $\alpha \geq 0$ represents the ARCH effects, $\beta \geq 0$ denotes the GARCH effects. The condition$\alpha + \beta < 1$ is necessary for the process to be covariance stationary. This ensures that the conditional variance converges to a finite long run mean and avoids explosive volatility dynamics. The three parameters ω, α, and β in the GARCH(1,1) volatility specification are obtained through the maximum likelihood estimation procedure.

2.2.2 Parameter Estimation

The parameter estimates of the Poisson and NB-INGARCH models were obtained through the conditional maximum likelihood estimation (CMLE) procedure using the *tsglm*() function from the *tscount* package in R software. The CMLE procedure for the INGARCH model operates similarly to the one utilized in the traditional GARCH model [11]. That is, the parameter values that maximizes the conditional likelihood function associated specified probability mass function are considered the optimal estimates of the model. The conditional likelihood function of the n observation $Y_1, Y_2, \ldots, Y_n$ conditionally on the pre-sample values is mathematically expressed as

$$L(\Theta) = \prod_{t=1}^{n} P(Y_t = y_t \,|\, \mathscr{F}_{t-1}), \tag{12}$$

where $P(Y_t = y_t \,|\, \mathscr{F}_{t-1})$ denotes the specified mass function conditional on the past information set and Θ represents the set of all parameters of interest in equation (8). The optimal values of Θ are obtained by maximizing the conditional log-likelihood which is expressed as

$$\ell(\Theta) = \log\left[L(\Theta)\right] = \sum_{t=1}^{n} \log\Big(P(Y_t = y_t \,|\, \mathscr{F}_{t-1}; \Theta)\Big). \tag{13}$$

The conditional log-likelihood function of the NB distribution has the form

$$\ell(\Theta) = \log\left[L(\Theta)\right] = \sum_{t=1}^{n} \left[Y_t \log(\lambda_t) - (r+Y_t)\log\big(1+\lambda_t\big) + \sum_{\nu=1}^{Y_t} \log(\nu + r - 1) - \log(Y_t!)\right],$$

where $\lambda_t = \omega + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} + f(t)$.

The parameter estimates $\hat{\Theta}$of the Θ and their corresponding asymptotic standard errors are obtained from the first and second derivatives of the conditional log-likelihood function with respect to the parameter of interest, $\theta_i, \quad i = 0, 1, \ldots, p+q+2k$. For any parameter θ_i, the first derivative has the form

$$\frac{\partial L(\Theta)}{\partial \theta_i} = \left(\frac{Y_t}{\lambda_t} - \frac{r+Y_t}{1+\lambda_t}\right) \cdot \frac{\partial \lambda_t}{\partial \theta_i}. \tag{14}$$

The second derivatives for θ_i, θ_j are

$$\frac{\partial^2 L(\Theta)}{\partial \theta_i \partial \theta_j} = \left(\frac{Y_t}{\lambda_t} - \frac{r+Y_t}{1+\lambda_t}\right) \cdot \frac{\partial^2 \lambda_t}{\partial \theta_i \partial \theta_j} - \left(\frac{Y_t}{\lambda_t^2} - \frac{r+Y_t}{(1+\lambda_t)^2}\right) \cdot \frac{\partial \lambda_t}{\partial \theta_i} \cdot \frac{\partial \lambda_t}{\partial \theta_j}, \tag{15}$$

for $i, j = 0, 1, \ldots, p+q+2k$. Furthermore, we have

$$\frac{\partial \lambda_t}{\partial \omega} = 1 + \sum_{m=1}^{q} \beta_m \frac{\partial \lambda_{t-m}}{\partial \omega},$$

$$\frac{\partial \lambda_t}{\partial \alpha_i} = Y_{t-i} + \sum_{m=1}^{q} \beta_m \frac{\partial \lambda_{t-m}}{\partial \alpha_i}, \quad i = 1, 2, \ldots, p,$$

$$\frac{\partial \lambda_t}{\partial \beta_j} = \lambda_{t-j} + \sum_{m=1}^{q} \beta_m \frac{\partial \lambda_{t-m}}{\partial \beta_j}, \quad j = 1, 2, \ldots, q,$$

$$\frac{\partial \lambda_t}{\partial \gamma_k} = \sin\left(\tfrac{2\pi kt}{S}\right) + \sum_{j=1}^{q} \beta_j \frac{\partial \lambda_{t-j}}{\partial \gamma_k}, \quad k = 1, 2, \ldots, K,$$

$$\frac{\partial \lambda_t}{\partial \delta_k} = \cos\left(\tfrac{2\pi kt}{S}\right) + \sum_{j=1}^{q} \beta_j \frac{\partial \lambda_{t-j}}{\partial \delta_k}, \quad k = 1, 2, \ldots, K.$$

For a large enough sample size n, the maximum likelihood estimator $\hat{\Theta}$ is consistent and approximated by the following distribution:

$$\hat{\Theta} \sim N\left(\Theta_0, \frac{1}{n} I(\Theta_0)^{-1}\right),$$

where $I(\Theta_0)$ denotes the information matrix at Θ_0, which is an array of actual values of the parameter vector Θ. The information matrix can be obtained by taking the expectation on both sides of equation (15) as follows:

$$E\left[\frac{\partial^2 L(\Theta)}{\partial\theta_i \partial\theta_j} \,\middle|\, \mathscr{F}_{t-1}\right] = \left(\frac{E(Y_t \mid \mathscr{F}_{t-1})}{\lambda_t} - \frac{r+E(Y_t \mid \mathscr{F}_{t-1})}{1+\lambda_t}\right) \cdot \frac{\partial^2 \lambda_t}{\partial\theta_i \partial\theta_j} - \left(\frac{E(Y_t \mid \mathscr{F}_{t-1})}{\lambda_t^2} - \frac{r+E(Y_t \mid \mathscr{F}_{t-1})}{(1+\lambda_t)^2}\right) \cdot \frac{\partial \lambda_t}{\partial\theta_i} \cdot \frac{\partial \lambda_t}{\partial\theta_j}.$$

Substituting $E(Y_t \mid \mathscr{F}_{t-1})$ by the results in equation (6) lead to

$$\left[\frac{\partial^2 L(\Theta)}{\partial\theta_i \partial\theta_j} \,\middle|\, \mathscr{F}_{t-1}\right] = -r\left(\frac{1}{\lambda_t} - \frac{1}{1+\lambda_t}\right) \cdot \frac{\partial \lambda_t}{\partial\theta_i} \cdot \frac{\partial \lambda_t}{\partial\theta_j}. \tag{16}$$

Using a similar approach, we obtain

$$E\left[\frac{\partial L(\Theta)}{\partial\theta_i}\frac{\partial L(\Theta)}{\partial\theta_j} \,\middle|\, \mathscr{F}_{t-1}\right] = E\left[\left(\frac{Y_t}{\lambda_t} - \frac{r+Y_t}{1+\lambda_t}\right)^2 \,\middle|\, \mathscr{F}_{t-1}\right] \cdot \frac{\partial \lambda_t}{\partial\theta_i} \cdot \frac{\partial \lambda_t}{\partial\theta_j} = r\left(\frac{1}{\lambda_t} - \frac{1}{1+\lambda_t}\right) \cdot \frac{\partial \lambda_t}{\partial\theta_i} \cdot \frac{\partial \lambda_t}{\partial\theta_j}. \tag{17}$$

The results from equations (16) and (18) lead to the following information matrix equality

$$-E\left[\frac{\partial^2 L(\Theta)}{\partial\theta_i \partial\theta_j} \,\middle|\, \mathscr{F}_{t-1}\right] = E\left[\frac{\partial L(\Theta)}{\partial\theta_i}\frac{\partial L(\Theta)}{\partial\theta_j} \,\middle|\, \mathscr{F}_{t-1}\right], \quad i,j = 0,1,\ldots,p+q+2k.$$

The asymptotic standard errors of the maximum likelihood estimators can be determined from the covariance matrix given as

$$E\left[(\hat{\Theta}-\Theta_0)(\hat{\Theta}-\Theta_0)^T\right] \equiv \frac{1}{n}\left(\hat{H}_n \hat{S}_n^{-1} \hat{H}_n\right)^{-1}, \tag{18}$$

where $\hat{H}_n = -\frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 L(\Theta)}{\partial\theta\partial\theta}$, and $\hat{S}_n^{-1} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial L(\Theta)}{\partial\theta}\left(\frac{\partial L(\Theta)}{\partial\theta}\right)^T$.

2.2.3 Model diagnostic

Once a time series model has been fitted and its parameters are estimated, it is then essential to evaluate the model's adequacy through various diagnostic checks to verify that the underlying assumptions are satisfied and that the fitted model provides a reliable representation of the data [30]. The study will utilise residual analysis, autocorrelation function (ACF) and partial ACF (PACF) of residuals, probability integral transformation (PIT), and AIC procedures to evaluate and compare fitted INGARCH processes.

Pearson residual analysis provides a standardized metric for assessing the difference between observed and expected values [31]. The Pearson residual is given by

$$\varepsilon_t = \begin{cases} \dfrac{Y_t - \hat{\lambda}_t}{\sqrt{\hat{\lambda}_t}}, \text{for Poisson-INGARCH} \\ \dfrac{Y_t - \hat{\lambda}_t}{\sqrt{\hat{\lambda}_t + \hat{r}\hat{\lambda}_t^2}}, \text{for NB-INGARCH,} \end{cases} \tag{19}$$

where Y_t is the observed time series, $\hat{\lambda}_t$ is the fitted conditional mean from the estimated INGARCH process, and $\hat{r}$ is the estimated dispersion parameter. The Pearson residuals of the correctly specified INAGARCH model should have a unit variance with a mean of zero and their residual plot over time should not depict any systematic pattern.

The ACF of INGARCH residuals quantifies the correlation between residuals at time t and their lagged values at time $t-k$ [32]. The ACF of lag k for residuals $\{\varepsilon_t\}$ is given by

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n} (\varepsilon_t - \bar{\varepsilon})(\varepsilon_{t-k} - \bar{\varepsilon})}{\sum_{t=1}^{n} (\varepsilon_t - \bar{\varepsilon})^2}, \tag{20}$$

with the mean of residuals $\bar{\varepsilon} = \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t$. When the ACF plot depicts significant spikes at certain lags (bars outside confidence bounds) it indicates that the residuals are correlated or that there is unexplained dependence at those lags. Therefore, insignificant residual autocorrelations mean the model is correctly specified.

The PIT histogram is an important diagnostic tool in count time series analysis for checking whether the fitted count model resembles the full conditional distribution of the data. The PIT value at time t, denoted by u_t, is defined as the conditional cumulative distribution function (CDF) of the count time series model evaluated at the observed outcome Y_t. This CDF is mathematically expressed as

$$u_t = F(Y_t \mid \mathscr{F}_{t-1}) = P(Y \leq Y_t \mid \mathscr{F}_{t-1}) = \sum_{k=0}^{y_t} f(k \mid \mathscr{F}_{t-1}), \tag{21}$$

where $f(k \mid \mathscr{F}_{t-1})$ is the conditional probability mass function. A count time series model is said to be correctly specified when the PIT values are uniformly distributed in the interval $[0,1]$. That is, an INGARCH model is said to be well-specified when the histogram of u_t is approximately uniform.

In statistical analysis, AIC is an information criterion established by [33] in 1973 for model selection. It evaluates competing models based on their relative quality, measuring performance by striking a balance between goodness of fit, parsimony, and sensitivity to parameter estimation. The formula for calculating AIC is given by

$$\text{AIC} = 2\phi - 2L(\hat{\Theta}), \tag{22}$$

where $L(\hat{\Theta})$ is the maximized log-likelihood of the fitted INGARCH model and ϕ is the dimension of the parameter vector Θ. The model with the smallest AIC value is considered a better trade-off between fitness and parsimony.

The mean absolute error (MAE) and root mean square error (RMSE) accuracy measures are utilized to evaluate the forecasting performance of the fitted models. The corresponding computing formulas are expressed as follows

$$\begin{cases} MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Y_t - \hat{Y}_t\right| \\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(Y_t - \hat{Y}_t\right)^2}, \end{cases} \tag{23}$$

where Y_t and $\hat{Y}_t$ are the observed and forecasted value of the series at time t, respectively. In comparative forecasting, a model with the lowest accuracy measure is typically regarded as the most reliable and thus considered the best performing specification.

3 Results

In this section, both Poisson and NB-INARCH(1) as well as INGARCH(1,1) were fitted to the four different datasets. The better fitting model is identified, and its residuals behavior is examined. INGARCH models incorporating Fourier terms are further applied to the series that exhibited weekly seasonal patterns.

Table 2 summarises the maximum likelihood estimates of the Poisson and NB-INARCH(1) and INGARCH(1,1) with their corresponding standard errors in parentheses. The AIC values for both model specifications are also provided. The parameter estimate of the model intercept is denoted by ω, the autoregressive term by α_1, the lagged conditional mean by β_1 and that of the dispersion parameter by r. These model parameters are statistically significant at 5% level, and meet the conditions, $\omega > 0$ and $\alpha_1 + \beta_1 < 1$. Therefore, we can conclude that the fitted models are stationary.

The NB-INGARCH framework consistently yields much lower AIC value compared to Poisson-INGARCH model across all the four datasets. These significant differences of the AIC values highlight a huge improvement achieved by

Table 2: Maximum likelihood estimates for Poisson and NB-INARCH(1) and INGARCH(1,1) models, with corresponding standard errors in parentheses.

Panel A: INARCH(1) Parameter Estimates and AIC values

	Poisson-INGARCH(1)				NB-INGARCH(1)				
Region	ω	α_1	β_1	AIC	r	ω	α_1	β_1	AIC
NAT	13.9960 (0.2918)	0.8850 (0.0036)	–	36618.52	0.7230	13.9960 (1.9801)	0.8850 (0.0526)	–	8796.731
KZN	4.5210 (0.1082)	0.7680 (0.0080)	–	18522.68	2.6530	4.5210 (0.6120)	0.7680 (0.1180)	–	5769.384
GP	5.1600 (0.1272)	0.7960 (0.0073)	–	17493.73	3.7650	5.1600 (0.8890)	0.7960 (0.1240)	–	6661.941
WC	5.2460 (0.1325)	0.8040 (0.0072)	–	17397.48	7.6270	5.2460 (1.4400)	0.8040 (0.1830)	–	7435.123

Panel B: INGARCH(1,1) Parameter Estimates and AIC values

	Poisson-INGARCH(1,1)				NB-INGARCH(1,1)				
Region	ω	α_1	β_1	AIC	r	ω	α_1	β_1	AIC
NAT	0.0004 (0.0015)	0.2457 (0.0041)	0.7541 (0.0041)	29945.42	0.5517	0.0004 (0.0169)	0.2457 (0.0432)	0.7541 (0.0410)	8549.221
KZN	0.0727 (0.0145)	0.2074 (0.0058)	0.7827 (0.0060)	14294.06	1.4373	0.0727 (0.0778)	0.2074 (0.0505)	0.7827 (0.0483)	5413.492
GP	0.0004 (0.0008)	0.1953 (0.0059)	0.8040 (0.0057)	313164.10	2.3312	0.0004 (0.0087)	0.1953 (0.0633)	0.8040 (0.0585)	6260.115
WC	0.2660 (0.0297)	0.2170 (0.0068)	0.7660 (0.0073)	13795.05	2.7940	0.2660 (0.2492)	0.2170 (0.0833)	0.7660 (0.0858)	6581.369

Note: NAT = National; KZN = KwaZulu-Natal; GP = Gauteng; WC = Western Cape.

incorporating the NB distribution, which effectively accommodates the overdispersion inherent in epidemiological count data. Moreover, the lagged conditional mean term enhances the model's performance since the AIC of NB-INARCH(1,1) is lower than that of NB-INARCH(1). It is worth noting that the lagged conditional mean term assumes relatively large values across all datasets, reflecting strong persistence in the conditional mean. These results suggest that the NB-INGARCH(1,1) model for COVID-19 death counts in South Africa and its three populous provinces relies heavily on the past expectations than the past realisations.

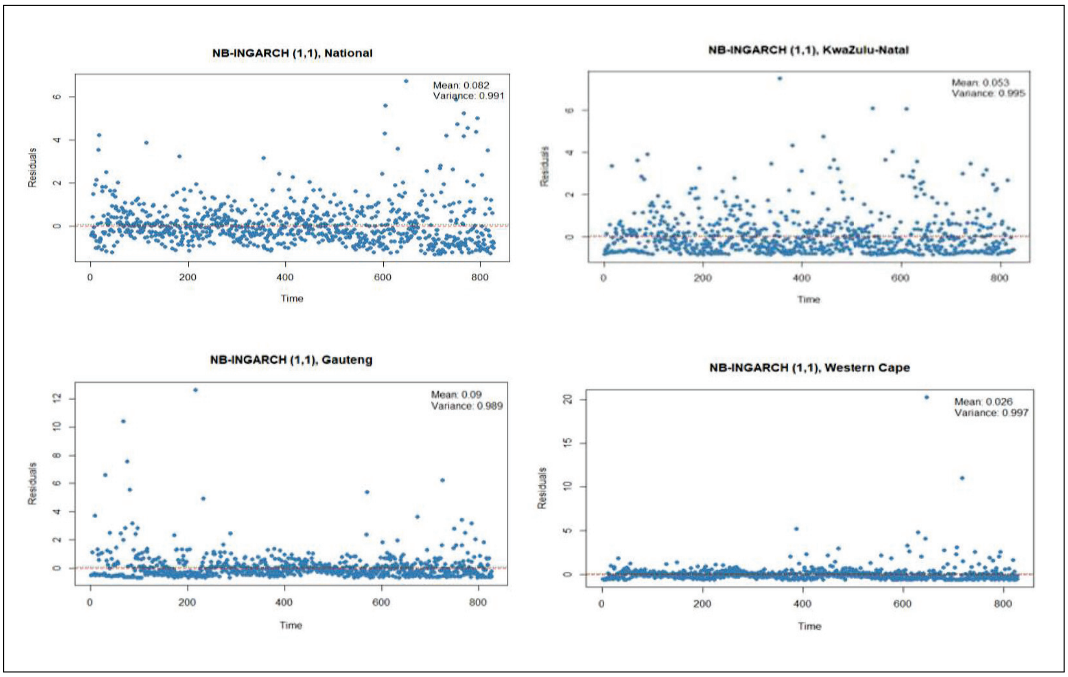

Fig. 3: Standardised Pearson residuals from the NB-INGARCH(1,1) models

The AIC results recommend the NB-INGARCH(1,1) framework as the better fit for all the datasets. Figure 3 depicts the standardised Pearson residuals of the NB-INGARCH(1,1) models. The mean and variance of these residuals are displayed on the top right corner of each residual plot. The plots of residuals against time does not depict any clear patterns which suggest that the variance of the residuals is constant over time. Moreover, the residuals exhibit characteristics of white noise, with their mean and variance approaching 0 and 1, respectively. However, there are few values above three standard deviations indicating the presence of outliers in the data. These outlier observations are not discarded from the data because [19] argued that NB-INGARCH framework can simultaneously capture overdispersion and extreme observations.

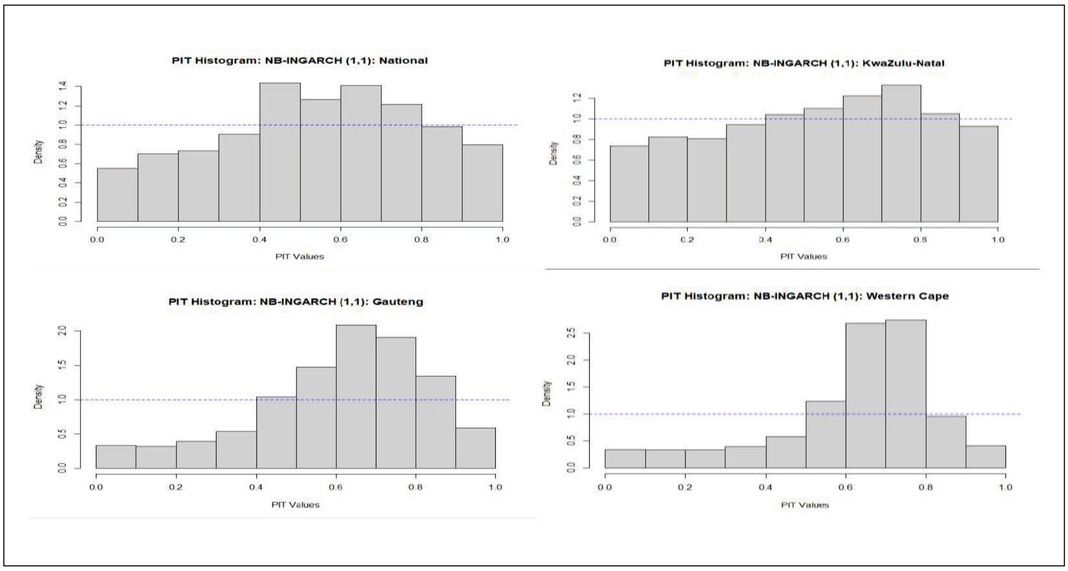

Fig. 4: PIT Histogram for the INGARCH(1,1) model across all the datasets

Figure 4 depicts the PIT histogram (densities against PIT values) of the NB-INGARCH(1,1) model for all the four datasets. For national and KwaZulu-Natal datasets, the plots show PIT values that approximate uniformity, reflecting well-calibrated forecasts and effective modeling of temporal structure. However, the Gauteng and Western Cape results reveal marked departures from uniformity, suggesting that these regions may benefit from more adaptive or region-specific modeling approaches.

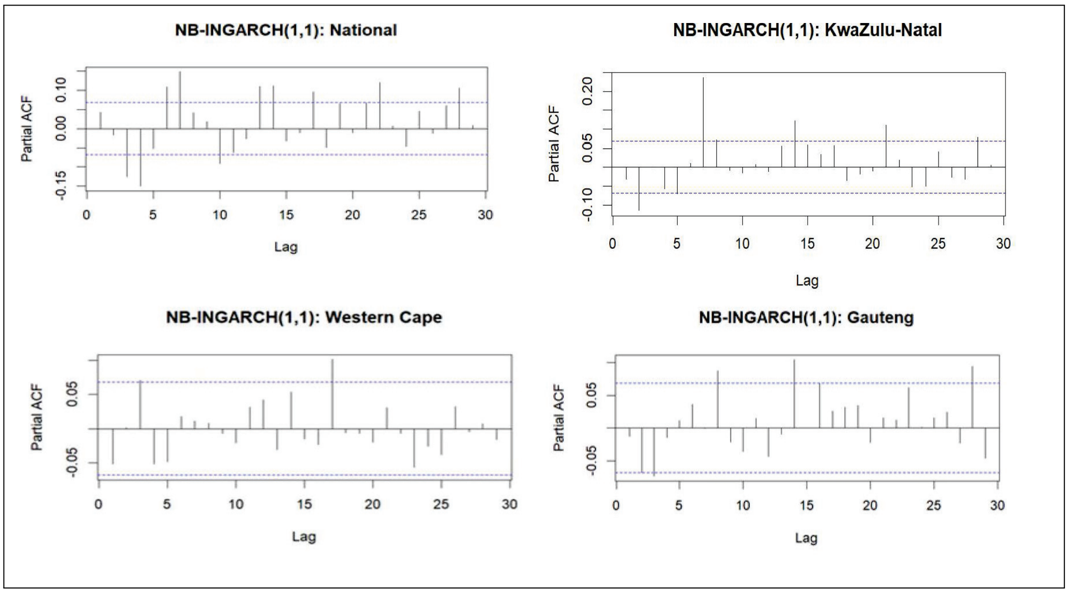

Fig. 5: PACF plots of standardised Pearson residuals of the NB-INGARCH(1,1) across all regions

Figure 5 depicts the PACF plots of the standardised Pearson residuals from the NB-INGARCH(1,1) models. For Gauteng and Western cape models, majority of the spikes between lag 2 and 30 are within the 95% confidence bounds. The Box-Pierce test statistic based on 10 lags are presented in Table 3. The p-values are greater than 5% level of significant indicating that NB-INGARCH(1,1) specifications provide an adequate fit to the Gauteng and Western cape data in terms of capturing serial autocorrelation within the residuals.

For national and KwaZulu-Natal model, the residuals exhibit significant correlation at lags 7, 14,21, and 28. This suggests a weekly cycle in the residuals, likely to reflect reporting practices or epidemic wave periodicity. The p-value of the Box-Pierce test statistic in Table 3 reject the null hypothesis of no serial autocorrelation within the residuals. The Box-Pierce results led to the conclusion that the NB-INGARCH(1,1) specification is not adequate to account for serial autocorrelation and weekly periodic effects in the national and KwaZulu-Natal count series. They emphasize the significance of incorporating the Fourier harmonic components in the NB-INGARCH(1,1) model.

Table 3: Box–Pierce test for serial autocorrelation within the model residuals. Null Hypothesis: No serial autocorrelation.

Region	Chi-squared Statistic	Degrees of freedom	p-value
National	101.3800	10	< 0.0001
KwaZulu-Natal	72.0210	10	< 0.0001
Gauteng	17.5400	10	0.0632
Western Cape	11.4740	10	0.3218

In conclusion, NB INGARCH(1,1) modelling framework demonstrates a superior and more suitable fit for the Gauteng and Western Cape series compared to Poison-INGARCH(1), Poisson-INGARCH(1,1) and NB-INGARCH(1) models. However, it's inability to account for serial autocorrelation and weekly periodic effects in the national and KwaZulu-Natal series renders it unsuitable for these datasets. The model's adequacy for Gauteng and Western Cape series stems from its ability to capture short term serial autocorrelation and overdispersion, with residuals resembling white noise. In contrast, the significant ACF and PACF lags at 7, 14, 21, and 28 supported by Box-Pierce test highlight the shortcoming of the NB INGARCH(1,1) specification to capture serial autocorrelation and weekly seasonality in the national and KwaZulu-Natal series. To address these limitations, Fourier harmonic components are incorporated into the NB INGARCH(1,1) framework.

The autocorrelation significant lags at 7, 14, 21, and 28 suggest the presence of weekly seasonality, thus the number of observations per cycle is $S = 7$. Accordingly, the NB INGARCH(1,1) framework with $k = 1, 2$, and 3 Fourier harmonic components were fitted to both the national and KwaZulu-Natal series. Table 4 presents the AIC values for the NB INGARCH(1,1) framework with $k = 1, 2$, and 3 Fourier harmonic components. The estimation results revealed that the optimal specification involves two Fourier harmonics ($K = 2$) for the national series, as indicated by the lowest AIC. In contrast, three harmonics terms ($K = 3$) provided the superior fit for the KwaZulu Natal series, again corresponding to the minimum AIC value. Table 5 Fourier presents the maximum likelihood estimates of the selected model specifications.

Table 4: AIC values corresponding to each Fourier harmonic component

Fourier harmonic terms, K	National	KwaZulu-Natal
	AIC	AIC
1	8446.416	5404.642
2	8444.117	5347.057
3	8462.550	5303.501

In the national series, the NB INGARCH(1,1) model with two Fourier harmonics yields parameter estimates whose 95% confidence intervals exclude zero, confirming statistical significance at the 5% level for all parameters. By contrast, in the KwaZulu Natal series, the NB INGARCH(1,1) specification with three harmonics produces confidence intervals for all cosine harmonic terms and the third sine harmonic that encompass zero, indicating insignificance at the 5% level. Despite these differences, model selection based on AIC shows that the NB INGARCH(1,1) with $K = 2$ harmonics achieves the lowest AIC value of 8444.117 for the national series, compared to the specification without harmonic terms.

Table 5: Parameter estimates of the NB-INGARCH(1,1) with Fourier harmonic terms

	National				KwaZulu-Natal			
Parameter	Est	SE	LCL	UCL	Est	SE	LCL	UCL
ω	0.1330	0.0533	0.0288	0.2380	0.1016	0.0438	0.0157	0.1870
α_1	0.7150	0.0474	0.6217	0.8080	0.7287	0.0536	0.6336	0.8440
β_1	0.2600	0.0440	0.1741	0.3470	0.2345	0.0499	0.1367	0.3320
γ_1	0.2730	0.0384	0.1975	0.3480	0.4171	0.0711	0.2777	0.5570
δ_1	0.1520	0.0383	0.0771	0.2270	0.0312	0.0692	-0.1044	0.1670
γ_2	0.1720	0.0661	0.0426	0.3020	0.5167	0.1205	0.2804	0.7530
δ_2	0.1480	0.0665	0.0180	0.2790	0.0506	0.1206	-0.1857	0.2870
γ_3	–	–	–	–	0.2606	0.1495	-0.0325	0.5540
δ_3	–	–	–	–	-0.1274	0.1490	-0.4196	0.1650
National: $r = 0.4940$, AIC = 8444.117					KwaZulu-Natal: $r = 1.0605$, AIC = 5303.501			

Note: Est = Estimate; SE = Standard error; LCL = Lower confidence limit; UCL = Upper confidence limit.

Similarly, the KwaZulu Natal series attains its minimum AIC value of 5404.642 when three harmonics are incorporated, improving upon the value of 5413.492 obtained from the NB-INGARCH(1,1) specification without Fourier components. The lower AIC values strengthen the hypothesis that incorporating Fourier components in the INGARCH framework enhances the epidemiological fit of the model specification. It worth noting that the serial autocorrelation and persistent weekly seasonality remains, as illustrated in Figure 6. The ACF diagnostics, reinforced by Box-Pierce test outcomes with p values below the 5% significance threshold, indicate that these specifications do not provide an adequate representation of the data.

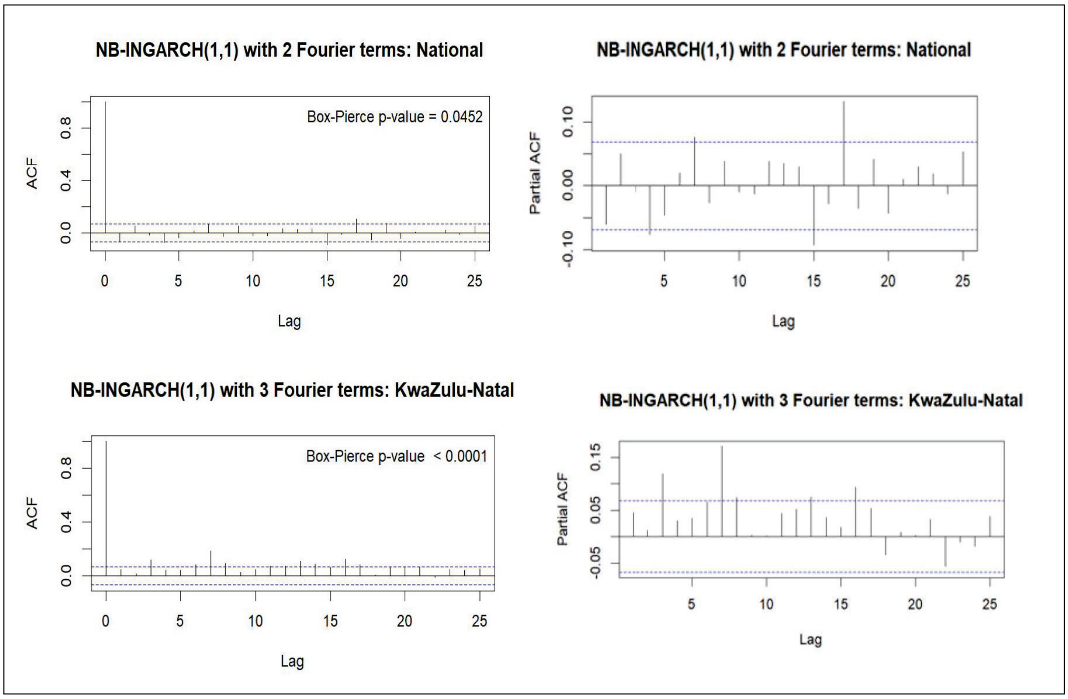

Fig. 6: PACF and ACF plot with p-value of the Box-Pierce test of the model residuals

The persistence of weekly autocorrelation, particularly evident at lag 7, motivated the inclusion of the seventh-lag autocorrelation term in the model. Consequently, the INGARCH(2,1) specification augmented with two Fourier harmonics was fitted to the national series, while the KwaZulu Natal series was modelled using three harmonics. The maximum likelihood estimates together with their corresponding standard errors, and 95% confidence intervals for the specified model are presented in Table 6.

Table 6: Parameter estimates of the NB-INGARCH(2,1) with Fourier harmonic terms

	National				KwaZulu-Natal			
Parameter	Est	SE	LCL	UCL	Est	SE	LCL	UCL
ω	0.1762	0.0720	0.0352	0.3170	0.1995	0.0973	0.0088	0.3900
α_1	0.2828	0.0498	0.1852	0.3810	0.2840	0.0679	0.1508	0.4170
α_7	0.0536	0.0535	-0.0512	0.1580	0.2126	0.0792	0.0575	0.3680
β_1	0.6309	0.0888	0.4569	0.8050	0.4526	0.1146	0.2279	0.6770
γ_1	0.2600	0.0397	0.1822	0.3380	0.3858	0.0795	0.2299	0.5420
δ_1	0.1373	0.0385	0.0618	0.2130	0.0206	0.0745	-0.1254	0.1670
γ_2	0.1665	0.0643	0.0406	0.2920	0.4942	0.1206	0.2578	0.7300
δ_2	0.1441	0.0646	0.0176	0.2710	0.0416	0.1112	-0.1763	0.2600
γ_3	–	–	–	–	0.2210	0.1340	-0.0516	0.4740
δ_3	–	–	–	–	-0.0712	0.1322	-0.3304	0.1880
National: $r = 0.4904$, AIC = 8444.335					KwaZulu-Natal: $r = 1.1667$, AIC = 5306.725			

Note: Est = Estimate; SE = Standard error; LCL = Lower confidence limit; UCL = Upper confidence limit.

The parameter estimate for the seventh lag autocorrelation term presented in Table 6 is statistically insignificant at the national level, indicating that weekly dependence does not play a meaningful role in explaining fluctuations in COVID 19 death counts at the national level. Conversely, the same term is statistically significant in KwaZulu Natal series, suggesting that weekly cycles exert a stronger influence on the progression of COVID 19 deaths in this province. The conditional mean lag (β_1) is consistently significant at the 5% level across both national and KwaZulu-Natal series, highlighting the persistence of short term dependence in COVID 19 deaths, where current mortality is strongly linked to the immediately preceding period. Confidence intervals for all Fourier harmonic components in the national model exclude zero, confirming that seasonal patterns captured by these harmonics are important drivers of national COVID 19 death dynamics. In KwaZulu Natal, only the first and second sine harmonics are significant, while the third sine and all cosine harmonics are insignificant, implying that the seasonal structure of COVID 19 deaths in this province is comparatively limited and less complex than the national pattern.

The Durbin-Watson (DW) test for serial autocorrelation in Table 7 yielded the value of the test statistic close to 2 (2.1557 and 2.0404, respectively) for both National and KwaZulu-Natal models. These test statistic values supported by the non significant p values (0.9875 and 0.6155) suggest that the NB-INGARCH(2,1) model with Fourier harmonic terms sufficiently capture short-lag dependence and that its residuals behave as approximately uncorrelated at the first lag. In contrast, the Box-Pierce test based on 10 lags produced conflicting results with the test statistic of 17.7770 and p-value of 0.0588 for the national model, indicating no evidence of serial autocorrelation across the 10 lags. However, for KwaZulu-Natal, the Box-Pierce statistic value of 51.9760 with p-value less than 0.0001 were obtained. This suggests a strong serial autocorrelation across the 10 lags. The DW and Box-Pierce tests results give an implication that while the models are adequate for short-lag dependence, they are not sufficient for medium-term dependence, particularly in KwaZulu-Natal. The results underscore the fact that uniform seasonal adjustments may be insufficient, and that tailored, region-specific modification may be sometimes required to improve medium to long-term dependence capture and overall adequacy.

Table 7: Test for serial autocorrelation of the residuals

	National		KwaZulu-Natal	
Test	Statistic	p-value	Statistic	p-value
Durbin Watson	2.1557	0.9875	2.0404	0.6155
Box-Pierce	17.7770	0.0588	51.9760	< 0.0001

The ACF and PACF plots of residuals for the NB INGARCH(2,1) model with Fourier harmonic components, shown in Figure 7, demonstrate that the specification effectively captured weekly seasonality in both the national and KwaZulu Natal series. The NB INGARCH(2,1) with Fourier harmonic components modelling framework successfully accounted for autocorrelation at lags 7, 14, and 21, confirming its ability to represent the weekly cycles in COVID 19 death counts. Although a few non-seasonal residual autocorrelation spikes remain significant, indicating that some dependence is not fully explained, the overall results suggest that the INGARCH(2,1) specification substantially addresses the dominant weekly seasonal structure in the data while leaving minor unexplained correlations.

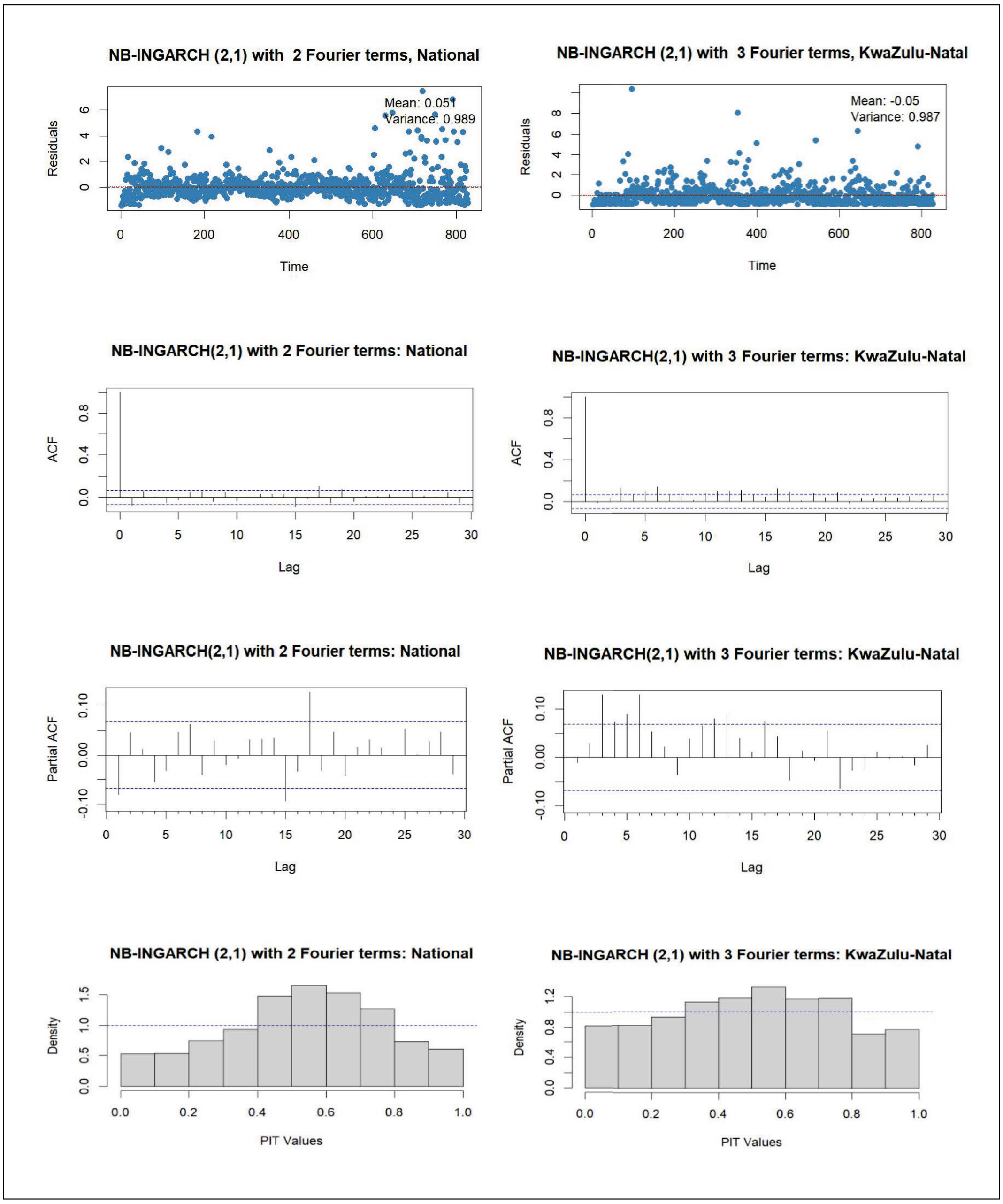

Fig. 7: Residual plot, ACF, PACF and PIT histogram for NB-INGARCH(2,1) model

The PIT histograms approximate a uniform distribution, implying that the model's predictive probabilities are well calibrated and provide reliable forecasts of the data. The standardised Pearson residuals of the NB INGARCH(2,1) model with the Fourier harmonics specifications are homoscedastic and tightly clustered around the mean of zero and approximately unit variance. This suggests that the chosen model specifications provide adequate fit to the data in terms of overdispersion, heteroscedasticity and weekly seasonality.

For comparative analysis, a ZINBA model incorporating the lagged death count (lag 1) and a time trend as predictor variables was fitted to both the national and KwaZulu-Natal series. These predictors are essential in capturing the dynamics of the process. The lag term reflects the influence of the previous day's mortality on current outcomes, while the trend variable accounts for the directional change in deaths over time. Following estimation, the standardized residuals from the ZINBA model were extracted and subsequently employed to estimate the parameters of the GARCH(1,1) process, the results of which are reported in Table 8.

Table 8: Maximum likelihood estimates with corresponding p-values for GARCH(1,1)

	National			KwaZulu-Natal		
Parameter	Est	SE	p-value	Est	SE	p-value
ω	0.0321	0.0124	0.0098	0.0139	0.0053	0.0083
α	0.0453	0.0171	0.0078	0.0910	0.0122	< 0.0001
β	0.9253	0.0227	< 0.0001	0.9080	0.0081	< 0.0001

Note: Est = Estimate; SE = Standard error.

The p-values for all parameter estimates in both the national and KwaZulu Natal series confirm statistical significance at the 5% level. In both cases, the GARCH effect β dominates, reflecting strong volatility persistence. For the national series, $\alpha+\beta = 0.9706$ satisfies the stationarity condition, indicating that shocks gradually disperse over time. By contrast, KwaZulu Natal's sum of $\alpha+\beta = 0.9990$ lies at the boundary of stationarity, suggesting prolonged volatility clustering likely driven by structural heterogeneity, localized shocks, or socio economic variability that weaken mean reversion.

Table 9: Forecasting accuracy measures

	National		KwaZulu-Natal	
Model	MAE	RMSE	MAE	RMSE
INGARCH(2,1) with Fourier terms	35.7333	58.9213	10.4343	20.3894
ZINBA-GARCH(1,1)	77.5891	173.9354	17.3431	36.6225

Table 9 reports the comparative results between the INGARCH(2,1) model with Fourier harmonic terms and the hybrid ZINBA–GARCH(1,1) specification. Across all evaluation metrics, the INGARCH(2,1) with Fourier harmonics (with $K = 2$ for the national series and $K = 3$ for KwaZulu Natal) consistently achieved the lowest values, indicating that this formulation provides superior forecasting performance relative to the ZINBA-GARCH(1,1) framework.

4 Summary and discussion

The search for the well-suited time series model for the daily number of COVID-19 death cases in South Africa and three of its populous provinces began with the Poisson and NB-INGARCH(1) model. Even though NB-INGARCH(1) model showed superiority performance over the Poisson-INGARCH(1) specifications, the PIT histogram suggested that the model is not well-specified. To address this shortcoming, the NB-INGARCH(1,1) was then explored across all the four datasets. The PIT histogram for the NB-INGARCH(1,1) model across all the four datasets are closer to the ideal uniform distribution compared to the PIT histogram of the NB-INGARCH(1) model. However, it didn't fully capture the weekly periodic effects in the national and KwaZulu-Natal data since their ACF plots of residuals depicted significant spikes at lags 7,14,21 and 28. Thus, the Fourier terms were incorporated in the NB-INGARCH(1,1) model to account for the weekly periodic effects in national and KwaZulu-Natal datasets. The NB-INGARCH(1,1) model with Fourier terms partially captured the weekly periodic structure in the national-level data and failed to moderately account for the periodic structure in the KwaZulu-Natal data. We then explored the NB-INGARCH(2,1) model incorporating the seventh-lag autocorrelation term and Fourier harmonic components (with K=2 for the national series and K=3 for KwaZulu Natal) to capture weekly seasonality. The model fully captured the weekly seasonality in both national and KwaZulu-Natal series. However, few non-seasonal autocorrelation lags remained significant, particularly in KwaZulu-Natal. The forecasting performance of the NB INGARCH(2,1) model was evaluated against the hybrid ZINBA-GARCH(1,1) framework, using MAE and RMSE metrics. The NB INGARCH(2,1) consistently achieved lower values across both series, indicating a more appropriate and superior fit. In the KwaZulu Natal series, however, the hybrid ZINBA-GARCH(1,1) model revealed strong shock sensitivity and near nonstationary persistence. The presence of unexplained non seasonal autocorrelation, coupled with stronger shock responsiveness and borderline stationarity in the KwaZulu Natal series, highlights the significance of further empirical investigation into province specific temporal structures underlying COVID 19 mortality trends.

The demonstrated superiority of the NB-INGARCH modelling framework over the Poisson-INGARCH specifications aligns with the findings of [15]. In that study, the NB-INGARCH models consistently outperformed the Poisson-INGARCH modelling approach when applied to overdispersed transaction counts and veto counts time series. In [19], an analysis of overdispersed polio data showed that the NB-INARCH(1) model achieved a lower AIC than the NB-INGARCH(1,1), indicating a better fit. Conversely, the findings of this current study suggest the opposite: the inclusion of the lagged conditional mean term improves model fit, as evidenced by the lower AIC of the NB-INARCH(1,1) compared to the NB-INARCH(1). In addition, the higher-order NB-INGARCH(2,1) specification provide superior performance compared to the widely adopted NB-INGARCH(1,1) model, thereby contradicting the conclusions reported by [17] and related studies. As established in the studies of [21, 23, 29] Fourier harmonic terms effectively capture seasonal patterns in the data. Incorporating Fourier harmonic components in the model enhances forecasting accuracy. This conclusion is consistent with the remark made by [22]. The standardised Pearson residuals of the selected model exhibit a mean near zero and approximately unit variance, indicating the model's adequacy for the data, as highlighted by [33].

5 Conclusion

The comparative analysis of alternative INGARCH specifications highlights the importance of model refinement in capturing the complex dynamics of COVID-19 mortality data in South Africa. Transitioning from the NB-INGARCH(1) to the NB-INGARCH(1,1) framework demonstrated the benefit of integrating a moving average component alongside a more adaptable distributional assumption for handling overdispersed count time series. The results also demonstrated that the incorporation of Fourier harmonic components in the NB-INGARCH specifications enhances the model?s performance. The NB-INGARCH(2,1) model embedded with Fourier harmonic components outperformed the hybrid ZINBA-GARCH(1,1) and adequately accounted for overdispersion, short-term dependence, heteroscedasticity, and weekly seasonality. However, few autocorrelation lags remained significant particularly in KwaZulu-Natal, suggesting that uniform seasonal adjustments and autocorrelation terms may be insufficient, and that tailored, region-specific modifications are sometimes required. Therefore, applying a blanket modelling strategy to epidemiological count data overlooks the complexity and variability inherent in disease processes. Datasets that appear comparable may, in fact, demand different methodological treatments to account for dispersion, autocorrelation, or seasonal effects. Effective analysis therefore necessitates a tailored modelling framework that aligns with the unique statistical and epidemiological features of each dataset.

Acknowledgement

The authors wish to express their gratitude to the anonymous referee for the thorough review of the manuscript and for the insightful comments that contributed significantly to the improvement of this paper.

References

[1] K. Chinhamu and R. Chifurira, "Evaluating South Africa?s market risk using asymmetric power auto-regressive conditional heteroscedastic model under heavy-tailed distributions," *Journal of Economic and Financial Sciences*, **12**(1), 1–11 (2019).

[2] K. Naradh, K. Chinhamu, and R. Chifurira, "Estimating the value-at-risk of JSE indices and South African exchange rate with Generalized Pareto and stable distributions," *Investment Management and Financial Innovations*, **18**(3), 151–165 (2021).

[3] K. Chinhamu, R. Chifurira, and E. Ranganai, "Value-at-risk estimation of precious metal returns using long memory GARCH models with heavy-tailed distribution," *Journal of Statistics Applications & Probability*, **11**(1), 89–107 (2022).
[4] K. Basira, L. Dhliwayo, K. Chinhamu, R. Chifurira, and F. Matarise, "Estimation and prediction of commodity returns using long memory volatility models," *Risks*, **12**(5), 73 (2024).
[5] S. D. Subramoney, K. Chinhamu, and R. Chifurira, "Value at Risk long memory volatility models with heavy-tailed distributions for cryptocurrencies," *Frontiers in Applied Mathematics and Statistics*, **11**, 1567626 (2025).
[6] T. H. Chavalala, R. Chifurira, K. Chinhamu, and J. Majakwara, "Volatile Count Modelling of COVID-19 Mortality Data: A Zero-Inflated Overdispersed Time Series Framework," *Statistics, Optimization & Information Computing*, **15**(4), 3305–3319 (2026).
[7] J. C. Reboredo, J. R. Barba-Queiruga, J. Ojea-Ferreiro, and F. Reyes-Santias, "Forecasting emergency department arrivals using INGARCH models," *Health Economics Review*, **13**(1), 51 (2023).
[8] I. Maingo, T. Ravele, and C. Sigauke, "Volatility Modelling of the JSE Top40 Index: Assessing the GAS Framework Against GARCH and Hybrid GARCH?XGBoost," *Journal of Risk and Financial Management*, **18**(12), 679 (2025).
[9] J. Lee and E. Hwang, "A generalized regime-switching integer-valued GARCH (1, 1) model and its volatility forecasting," *Communications for Statistical Applications and Methods*, **25**(1), 29–42 (2018).
[10] L. Qian and F. Zhu, "A flexible model for time series of counts with overdispersion or underdispersion, zero-inflation and heavy-tailedness," *Communications in Mathematics and Statistics*, **13**(2), 431–454 (2025).
[11] R. Ferland, A. Latour, and D. Oraichi, "Integer-valued GARCH process," *Journal of Time Series Analysis*, **27**(6), 923–942 (2006).
[12] A. R. Jamaludin, F. Yusof, and S. Suhartono, "Modelling asthma cases by count analysis approach: Poisson INGARCH and negative binomial INGARCH," *Matematika*, 15–30 (2020).
[13] J. Chicheli, S. Jere, O. Hamweene, and E. Moyo, "Modelling Malaria Cases using Count Analysis Approach: COM-Poisson INGARCH and NB-INGARCH," *International Journal of Advanced Applied Mathematics and Mechanics (IJAAMM)*, **13**(1), 95–104 (2025).
[14] F. Zhu, "Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models," *Journal of Mathematical Analysis and Applications*, **389**(1), 58–71 (2012a).
[15] F. Zhu, "Modeling time series of counts with COM-Poisson INGARCH models," *Mathematical and Computer Modelling*, **56**(9–10), 191–203 (2012b).
[16] C. Mulamfu, "A negative binomial INGARCH model for overdispersed count time series structure, parameter estimation and application to real data," Doctoral dissertation, The University of Zambia (2025).
[17] E. Mubanga, S. Jere, and E. Moyo, "A comparative study of the Poisson and Negative Binomial INGARCH model: Application to malaria cases in Chembe District, Zambia," (2025).
[18] K. F. Sellers, S. Borle, and G. Shmueli, "The COM-Poisson model for count data: a survey of methods and applications," *Applied Stochastic Models in Business and Industry*, **28**(2), 104–116 (2012).
[19] F. Zhu, "A negative binomial integer-valued GARCH model," *Journal of Time Series Analysis*, **32**(1), 54–67 (2011).
[20] M. Awale, A. S. Kashikar, and T. V. Ramanathan, "Modeling seasonal epidemic data using integer autoregressive model based on binomial thinning," *Model Assisted Statistics and Applications*, **15**(1), 1–17 (2020).
[21] A. Norzaida, M. D. Zalina, and F. Yusof, "Application of Fourier series in managing the seasonality of convective and monsoon rainfall," *Hydrological Sciences Journal*, **61**(10), 1967–1980 (2016).
[22] C. W. Chen, C. S. Chen, and H. H. Liao, "Advances in spatial integer-valued time series modeling," *Journal of Applied Statistics*, 1–24 (2026).
[23] L. G. Letema and H. A. Mbwambo, "Forecasting Tourist Arrivals in Tanzania Using INGARCH Models: Accounting for Seasonality and COVID-19 Impacts," *International Journal of Management, Accounting & Economics*, **13**(2), 234 (2026).
[24] M. Liu, F. Zhu, J. Li, and C. Sun, "A systematic review of INGARCH models for integer-valued time series," *Entropy*, **25**(6), 922 (2023).
[25] J. Yang and L. Qian, "Parsimonious and Efficient Integer-valued Autoregressive Process: Inference and Application," *PREPRINT (Version 1)*, available at Research Square, https://doi.org/10.21203/rs.3.rs-9064129/v1 (2026).
[26] M. Yang, G. K. Zamba, and J. E. Cavanaugh, "Markov regression models for count time series with excess zeros: A partial likelihood approach," *Statistical Methodology*, **14**, 26–38 (2013).
[27] J. H. Jakobsen, "Application of count time series to battle deaths," Master's thesis, University of Oslo (2021).
[28] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, **31**(3), 307–327 (1986).
[29] N. A. Mustaffa, S. M. Zahari, N. A. Farhana, N. Nasir, and A. H. Azil, "Forecasting the incidence of dengue fever in Malaysia: A comparative analysis of seasonal ARIMA, dynamic harmonic regression, and neural network models," *International Journal of Advanced and Applied Sciences*, **11**(1), 20–31 (2024).
[30] R. J. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice," OTexts (2018).
[31] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, "Time series analysis: forecasting and control," John Wiley & Sons (2015).
[32] K. Fokianos and D. Tjøstheim, "Log-linear Poisson autoregression," *Journal of Multivariate Analysis*, **102**(3), 563–578 (2011).
[33] H. Akaike, "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, **60**(2), 255–265 (1973).