STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, April 2026, pp 3305–3319.
Published online in International Academic Press (www.IAPress.org)

Volatile Count Modelling of COVID-19 Mortality Data: A Zero-Inflated Overdispersed Time Series Framework

Thembhan Hlayisani Chavalala [1,*], Retius Chifurira [2], Knowledge Chinhamu [2], 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*

Abstract Epidemiological count time series often display challenging characteristics such as overdispersion, zero-inflation, and serial dependence. This study explores appropriate statistical frameworks for modelling such data, using daily COVID-19 mortality counts from South Africa and its three most populous provinces as a case study. The observed data exhibited strong serial autocorrelation, excess zeros, overdispersion, and time-varying volatility. To capture these dynamics, we employed hybrid models combining zero-inflated Poisson autoregressive (ZIPA) and zero-inflated negative binomial autoregressive (ZINBA) structures with a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) component. Model comparisons using the Vuong test indicated that the ZINBA model offered a superior fit. Further, a GARCH model applied to the ZINBA residuals effectively accounted for residual heteroscedasticity, as validated by sign-bias testing. However, the forecasting accuracy metrics consistently favored the simpler ZINBA model over the ZINBA-GARCH model. Therefore, model selection should be guided not only by statistical diagnostics but also by the intended application.

Keywords Zero-inflation, Overdispersion, Heteroscedasticity, Poisson, Negative Binomial, Autoregressive model, GARCH

AMS 2010 subject classifications 46N30, 62C10, 62F15

DOI: 10.19139/soic-2310-5070-2734

1. Introduction

Discrete count time series refers to a sequences of non-negative integer observations recorded over time. It arises in numerous disciplines, including epidemiology (e.g., daily counts of COVID-19-related deaths), criminology (e.g., monthly assault cases), and environmental science (e.g., annual flood occurrences). These data often display complex features such as zero-inflation (excess zero counts), overdispersion (variance exceeding the mean), underdispersion (variance less than the mean), autocorrelation (temporal dependence), and heteroscedasticity (time-varying variance). Traditional modelling approaches frequently rely on the Poisson distribution, which assumes equidispersion, a constraint that is often violated in real-world applications. As a result, models based solely on Poisson assumptions may yield biased or inefficient estimates. Recognizing and accurately characterizing the underlying distributional and temporal properties of count time series is therefore essential for selecting appropriate models and drawing reliable inferences. This is particularly critical in high-stakes contexts such as epidemiological surveillance and public health decision-making.

[*] Correspondence to: Thembhani Hlayisani Chavalala (Email: thembhani.chavalala@ul.ac.za). Department of Statistics and Operations Research, University of Limpopo, Limpopo Province, Polokwane, South Africa(0727).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

An overwhelming number of studies have utilised the Box and Jenkins (1970) methodology to model and forecast the number of COVID-19 deaths around the globe [see [1] - [4]]. The assumptions of the Box-Jenkins methodology render it inadequate for modelling the dynamics of the count data. Alzhrani (2022) [5] utilised the count log-linear Poisson autoregressive model and Box-Jenkins methodology to study the dynamic of COVID-19 in Saudi Arabia. The forecasting metrics revealed that the count log-linear Poisson autoregressive model provides better predictions when compared with the Box-Jenkins methodology. While the model could be providing a superior fit in some respect, it does not incorporate overdispersion, heteroscedasticity and zero-inflation which are some of the key characteristics of the count data.

Mthethwa et al., (2022) [6] utilised the Markov switching GARCH-type model combined with heavy-tailed distribution to estimate the minimum daily COVID-19 deaths in South Africa. The standard GARCH-type models are tailored for real-valued continuous time series data, particularly the one that displays volatility clustering over time. However, these models are not appropriate for modelling count data that exhibits some form of serial dependence [7]. The number of deaths for COVID-19 are integer-valued time series and according to [8] various techniques for modelling this type of data have been proposed and studied. Tawiah et al., (2021) [9] studied the counts of COVID-19 mortalities in Ghana using the zero-inflated time series modelling techniques. Nevertheless, these techniques do not incorporate the volatility clustering exhibited by the daily deaths due to COVID-19 studied by [6] . Therefore, the current study seeks to employ hybrid models combining the zero-inflated time series modelling structures considered by [9] with the GARCH component studied by [6].

The primary objective of this study was to explore the discrete count time series models that can be utilised to model and forecast datasets showing the characteristics similar to those that were observed in the South African COVID-19 dataset. To achieve this goal, this study follows the systematic approach. First, exploratory data analysis was conducted to obtain a comprehensive understanding of the characteristics of the dataset. This was followed by the evaluation of appropriate count time series models and the selection of the most effective model. Finally, the selected model is employed for forecasting.

2. Data

This section provides the description and exploration of the four datasets considered. It begins by describing the data, then discussing the methods to be utilised for exploratory data analysis. Lastly, carry out the exploratory data analysis and suggest the suitable framework for modelling the data.

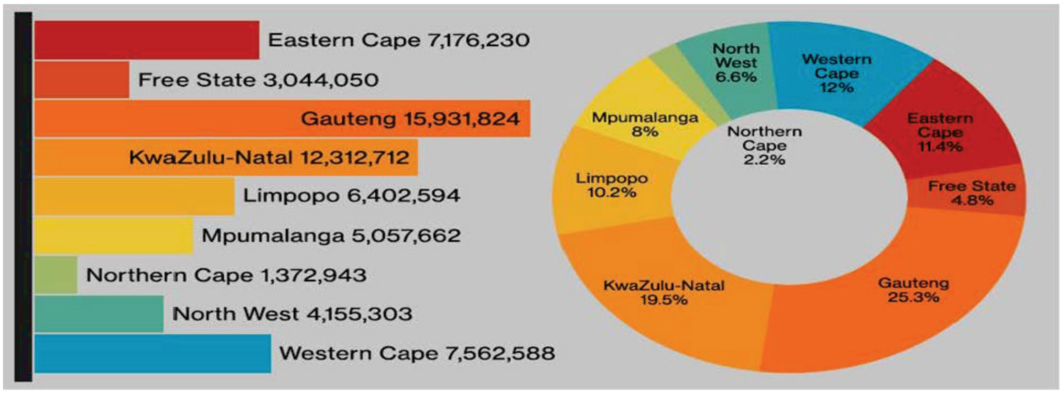

Figure 1. The population of South Africa's nine provinces (Alexander, 2024) [10].

2.1. Data description

This study explores the daily COVID-19 death cases recorded from March 27, 2020, to July 22, 2022. The dataset was published by the National Institute for Communicable Diseases (NICD) in South Africa and can be accessed on the website: https://github.com/dsfsi/covid19za/blob/master/data/.

Figure 1 portrays the demographic of the nine provinces of South Africa. In terms of population, the province of Gauteng is largest, followed by KwaZulu-Natal, then Western Cape. The current study focuses on the national dataset and three of its largest provinces as depicted by Figure 1. These three provinces were considered COVID-19 epicenters.

2.2. Exploratory data analysis

The most encountered characteristics of the discrete count time series data in the field of epidemiology are non-negative integers with large frequency of zeros, autocorrelated and overdispersed with heteroscedasticity ([11] and [12]). Therefore, this section discusses the statistical measures that will be used to establish the existence of these features in the considered datasets.

The degree of dispersion was assessed by the *index of dispersion*, which is the ratio of the variance to the mean of the dataset. That is, suppose that S^2 and $\bar{Y}$ are the variance and mean of the data at hand, respectively. Then, the index of dispersion is given by $\frac{S^2}{\bar{Y}}$. An index of dispersion value greater than one confirms that the data is overdispersed.

The *zero-inflation index* was employed to measure the excess of zeros in the count COVID-19 death cases in the datasets. An inflation index value greater than zero suggests that the dataset is zero-inflated and should be analysed using zero-inflated count time series models. The zero-inflation index is computed using the following mathematical expression:

$$1 + \frac{\log(\hat{\theta})}{\bar{Y}},$$

where $\bar{Y}$ is the mean of the dataset, and $\log(\hat{\theta})$ is the natural logarithm of the proportion of zeros in the dataset.

The existence of serial autocorrelation in the datasets was examined through the Durbin-Watson, Box-Pierce, and Box-Ljung tests. The time series plot and the ARCH Lagrange Multiplier (ARCH-LM) test were utilised to assess the variation in the variance over time (heteroscedasticity).

Figure 2 shows time series plots demonstrating the daily COVID-19 death counts for (a) South Africa at the national level and the three populous provinces: (b) KwaZulu-Natal, (c) Western Cape, and (d) Gauteng. The trends reveal noticeable fluctuations in mortality figures across all four datasets, highlighting periods of both high and low variability, which are suggestive of heteroscedasticity. Moreover, the presence of days with zero recorded deaths is evident across the datasets (as also shown in Table 1, where the minimum value for each series is zero). To assess the impact of these zero counts, the datasets were further analyzed by comparing means and variances, and through computing both the dispersion index and the zero-inflation index.

Table 1 depicts that in all the four datasets the mean is lower than the variance which leads to the dispersion index to be greater than one and suggests that there is overdispersion in these datasets. Around 10% of the national daily COVID-19 death counts are recorded as zeros. The three epicenters recorded the following relative frequency of the zero COVID-19 death count: 13% in Gauteng, 13% in KwaZulu Natal, and 12% in the Western Cape. The significance of these relative frequencies of the zero COVID-19 death count is that all the datasets are zero-inflated, as the zero-inflation index is greater than zero.

The datasets were tested for serial autocorrelation using Durbin-Watson, Box-Pierce, and Box-Ljung techniques. The p-values associated with the test statistics of all three techniques are presented in Table 2. All the p-values are less than the 5% level of significance, which suggests the rejection of the null hypothesis, leading to the conclusion that there is autocorrelation across the datasets.

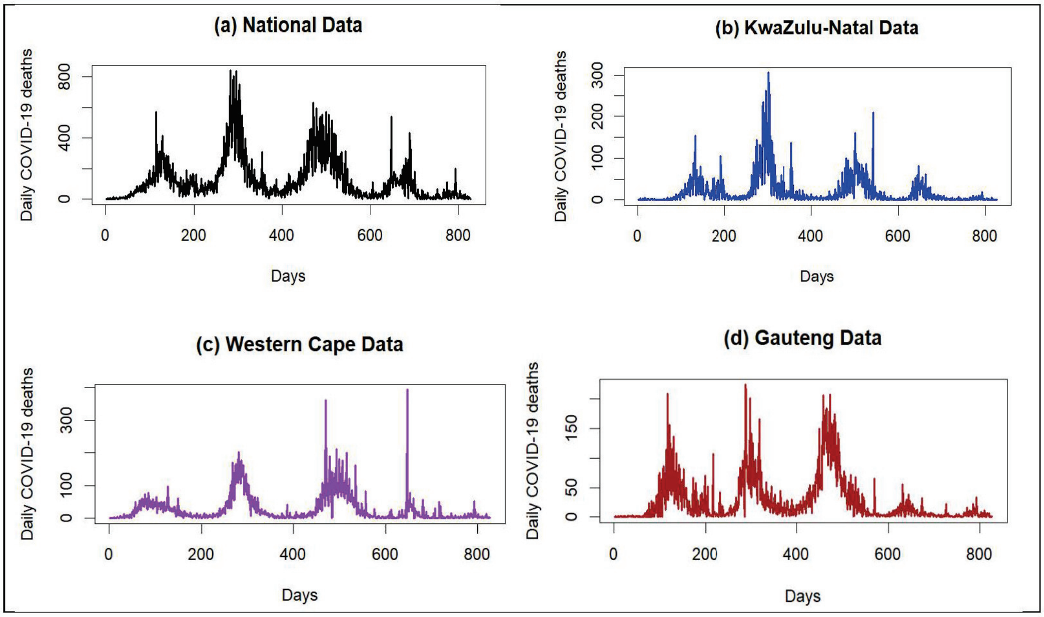

Figure 2. Time series plots of the daily COVID-19 deaths in South Africa and three largest Provinces .

Table 1. Descriptive statistics for daily death counts of COVID-19

Statistic	National	Gauteng	KwaZulu-Natal	Western Cape
Mean	123.2684	25.40508	19.66747	26.96372
Variance	20094.97	1415.304	1354.564	1660.919
Proportion of zeros	0.01451	0.13422	0.206771	0.11971
Dispersion index	163.0180	55.6701	68.8733	55.4310
Zero-inflation index	0.9657	0.9209	0.9199	0.9213
Minimum	0	0	0	0
Maximum	844	306	225	394

Table 2. Tests for Autocorrelation in the Daily COVID-19 Death Counts
Null Hypothesis: No Autocorrelation

Test	National		KwaZulu-Natal		Gauteng		Western Cape	
	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

Table 3. Test for ARCH Effects in the Daily COVID-19 Death Counts
Null Hypothesis: No ARCH Effects

Test	National		KwaZulu-Natal		Gauteng		Western Cape	
	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

Table 3 depicts that all p-values corresponding to the ARCH-LM test for national and the three epicenter Provinces are less than 0.01%, suggesting that the null hypothesis which states that there are no ARCH effects in the data should be rejected. Therefore, it can be concluded with 5% level of significance that there is inconstant variation in the variance of the datasets.

In summary, the daily number of deaths caused by COVID-19 for national and the three Provinces exhibits the following characteristics:

- Non-negative integers
- Overdispersion
- Zero-inflation
- Autocorrelation
- Volatility

According to [11] - [14] count time series models provide a better framework for modelling datasets with the characteristics listed above than the traditional and standard time series models. Therefore, this study explores the Zero-Inflated Poisson Autoregressive (ZIPA) model. Moreover, the ZIPA was extended to Zero-Inflated Negative Binomial Autoregressive (ZINBA) model. This extension was introduced to effectively account for the overdispersion that exists in the datasets. The residuals of the better fit between the ZIPA and ZINBA were utilised to fit the GARCH model which is most suitable for handling volatile datasets.

3. Methodology

This section discusses the zero-inflated count time series models, parameter estimation method, model comparison criterion, time series model for volatile dataset, and the asymmetric test for leverage effects.

3.1. Zero-Inflated Poisson Autoregressive Model

The zero-inflated Poisson (ZIP) model by [15] has been a useful tool in modelling the count time series with excess zeros, ([13], [9], [5], [16], and [17]). The ZIP model is built upon the foundation of the ZIP probability mass function (PMF). Let Y_t, $t = 1, 2, \ldots, N$ denote the response count variable. Then $Y_t \sim \text{ZIP}(\mu_t, \theta_t)$ if its probability mass function (PMF) is given by:

$$f_{Y_t}(y_t \mid \mathcal{F}_{t-1}) = \begin{cases} \theta_t + (1-\theta_t)e^{-\mu_t}, & \text{if } y_t = 0 \\ (1-\theta_t)\dfrac{\mu_t^{y_t} e^{-\mu_t}}{y_t!}, & \text{if } y_t > 0 \end{cases} \tag{1}$$

where $\mathcal{F}_{t-1}$ is the filtration of all known information at time $t-1$, including the first lag of the daily COVID-19 death counts and the trend variable. The parameter $\theta_t \in (0, 1)$ is the zero-inflation parameter, and μ_t is the intensity parameter. The expected value of the ZIP distribution is:

$$\mathbb{E}(Y_t \mid \mathcal{F}_{t-1}) = \sum_{y=0}^{\infty} y \cdot P(Y_t = y \mid \mathcal{F}_{t-1}) \tag{2}$$

For $y = 0$, the expectation is zero. For $y > 0$, we have:

$$\mathbb{E}(Y_t \mid \mathcal{F}_{t-1}) = (1-\theta_t)\sum_{y=1}^{\infty} \frac{y\mu_t^{y} e^{-\mu_t}}{y!} = \mu_t(1-\theta_t) \tag{3}$$

The variance of the ZIP distribution is derived using the law of total variance:

$$\text{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mathbb{E}[\text{Var}(Y_t \mid Z)] + \text{Var}[\mathbb{E}(Y_t \mid Z)] \tag{4}$$

Let Z be a latent indicator variable, where $Z = 1$ if the observation comes from the Poisson component and $Z = 0$ if it corresponds to a structural zero. Then:

$$\text{Var}(Y_t \mid Z = 1) = \mu_t, \quad \mathbb{E}[\text{Var}(Y_t \mid Z = 1)] = \mu_t(1 - \theta_t)$$

$$\mathbb{E}(Y_t \mid Z = 1) = \mu_t, \quad \mathbb{E}(Y_t \mid Z = 0) = 0$$

$$\text{Var}[\mathbb{E}(Y_t \mid Z)] = \theta_t(1 - \theta_t)\mu_t^2$$

Thus, the total variance is:

$$\text{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mu_t(1 - \theta_t) + \theta_t(1 - \theta_t)\mu_t^2 = \mu_t(1 - \theta_t)(1 + \mu_t\theta_t) \tag{5}$$

The dispersion index is:

$$\frac{\text{Var}(Y_t \mid \mathcal{F}_{t-1})}{\mathbb{E}(Y_t \mid \mathcal{F}_{t-1})} = \frac{\mu_t(1 - \theta_t)(1 + \mu_t\theta_t)}{\mu_t(1 - \theta_t)} = 1 + \mu_t\theta_t \tag{6}$$

When $\theta_t = 0$, the dispersion index equals 1, and the ZIP distribution reduces to the ordinary Poisson distribution.

The ZIP distribution is a mixture of two components: the intensity parameter and the zero-inflation component. The ZIPA technique models these separately as follows:

$$\log(\mu_t) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t = X^\top \beta \tag{7}$$

$$\text{logit}(\theta_t) = \log\left(\frac{\theta_t}{1 - \theta_t}\right) = \gamma_0 + \gamma_1 Y_{t-1} + \gamma_2 t = X^\top \gamma \tag{8}$$

Here, $X^\top = [1, Y_{t-1}, t]$ is the vector of explanatory variables. The vector $\beta = [\beta_0, \beta_1, \beta_2]^\top$ contains the coefficients for the log-linear Poisson intensity model, while $\gamma = [\gamma_0, \gamma_1, \gamma_2]^\top$ represents the coefficients of the logit model governing the zero-inflation probability.

3.2. Zero-inflated Negative Binomial Autoregressive Model

To simultaneously address the excess zeros and overdispersion observed in the datasets presented in Section 2, we extend the ZIPA model to a more general ZINBA model using the Zero-Inflated Negative Binomial (ZINB) distribution, which is a well-known technique for modeling overdispersed count data. The probability mass function (PMF) of the ZINB model is given by:

$$f_{Y_t}(y_t \mid \mathcal{F}_{t-1}) = \begin{cases} \theta_t + (1 - \theta_t)\left(\frac{k_t}{k_t + \mu_t}\right)^{k_t}, & \text{if } y_t = 0 \\ (1 - \theta_t) \cdot \frac{\Gamma(k_t + y_t)}{\Gamma(k_t) y_t!}\left(\frac{k_t}{k_t + \mu_t}\right)^{k_t}\left(\frac{\mu_t}{k_t + \mu_t}\right)^{y_t}, & \text{if } y_t > 0 \end{cases} \tag{9}$$

where k_t, θ_t, and μ_t are the dispersion, zero-inflation, and intensity parameters, respectively. Let Z be a latent indicator variable, taking value 1 when the observation comes from the negative binomial component, and 0 when it corresponds to a structural zero. The conditional mean of the ZINB distribution is:

$$\mathbb{E}(Y_t \mid \mathcal{F}_{t-1}) = \mathbb{E}[\mathbb{E}(Y_t \mid Z, \mathcal{F}_{t-1})] \tag{10}$$

When $Z = 0$, then $Y_t = 0$. When $Z = 1$, then $Y_t \sim \text{NB}(\mu_t, k_t)$, and:

$$\mathbb{E}(Y_t \mid \mathcal{F}_{t-1}) = \mu_t(1 - \theta_t) \tag{11}$$

The variance of the ZINB distribution is derived using the law of total variance:

$$\text{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mathbb{E}[\text{Var}(Y_t \mid Z)] + \text{Var}[\mathbb{E}(Y_t \mid Z)] \tag{12}$$

If $Z = 1$ and $Y_t \sim \text{NB}(\mu_t, k_t)$, then:

$$\text{Var}(Y_t \mid Z = 1) = \mu_t + \frac{\mu_t^2}{k_t}, \quad \mathbb{E}[\text{Var}(Y_t \mid Z)] = (1 - \theta_t)\left(\mu_t + \frac{\mu_t^2}{k_t}\right) \tag{13}$$

The variance of the conditional mean is:

$$\text{Var}[\mathbb{E}(Y_t \mid Z)] = \theta_t(1 - \theta_t)\mu_t^2 \tag{14}$$

Combining the two components yields:

$$\text{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mu_t(1 - \theta_t)\left(1 + \mu_t\theta_t + \frac{\mu_t}{k_t}\right) \tag{15}$$

The dispersion index of the ZINB distribution is:

$$\frac{\text{Var}(Y_t \mid \mathcal{F}_{t-1})}{\mathbb{E}(Y_t \mid \mathcal{F}_{t-1})} = 1 + \mu_t\theta_t + \frac{\mu_t}{k_t} \tag{16}$$

Like the ZIP distribution, the ZINB distribution is a mixture of the negative binomial with a degenerate distribution having point mass at zero. The ZINBA technique models the intensity parameter μ_t and the zero-inflation parameter θ_t as described in equations (7) and (8), respectively. It further models the dispersion parameter k_t as follows:

$$\log(k_t) = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 t = X^\top \alpha \tag{17}$$

where $X^\top = [1, Y_{t-1}, t]$ represents the vector of past explanatory variables, and $\alpha = [\alpha_0, \alpha_1, \alpha_2]^\top$ is the vector of regression coefficients.

3.3. GARCH Model

The generalised ARCH of order p and q [GARCH (p,q)] model by [18] is the extension of the ARCH which was first introduced by [19]. This study adopts the GARCH (1,1) model because of its straightforward structure and proven efficiency. Its limited number of parameters makes it easier to estimate and interpret, while also minimizing the likelihood of overfitting. Additionally, the model tends to yield more stable and reliable convergence during the estimation process. The mathematical formulation of the model is as follows:

$$\sigma_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2 \tag{18}$$

where $\beta_1, \alpha_1 \geq 0$ are model parameters, ϵ_{t-1}^2 is the ARCH component, σ_{t-1}^2 is the GARCH component, and ω is the intercept. The model is said to be *weakly stationary* when $\beta_1 + \alpha_1 < 1$.

The standard GARCH (1,1) model assumes that financial market and/or epidemiological time series respond symmetrically to both positive and negative shocks. However, in practice, these shocks often have asymmetric impacts on volatility. Therefore, it is important to evaluate whether the standard GARCH specification sufficiently captures this behavior, or if an asymmetric variant such as EGARCH, TGARCH, or GJR-GARCH is more appropriate. To perform this evaluation, the sign-bias test will be employed.

3.4. Sign-Bias Test

Engle and Ng (1993) [20] introduced a tool known as the sign-bias test to assess whether the financial market reacts differently to bad and good news. The value of the sign-bias test statistic is calculated by regressing the squared standardised residuals on lagged negative and positive shocks. It is defined as the t-ratio for the regression coefficient b and is mathematically expressed as:

$$z_t^2 = c_0 + bs_t^- + z_{t-1}\gamma + \delta_t \tag{19}$$

where s_t^- is a binary predictor variable that takes the value 1 if $\epsilon_{t-1} < 0$, and 0 otherwise. Here, z_{t-1} may represent other variables included in the regression model, and δ_t is the error term.

3.5. Parameter Estimation

The parameter estimates of both the ZIPA and ZINBA models are obtained through the $zeroinfl()$ function from the $pscl$ package in R, which performs Maximum Likelihood Estimation (MLE). Even though the theoretical framework follows the expectation maximization algorithm developed by [14], our application relies on the built-in optimization routine provided by $zeroinfl()$ function. This routine does not require manual specification of initial values and uses default convergence criteria based on the $optim()$ function.

This algorithm computes the maximum partial likelihood estimator and the corresponding standard error using the partial likelihood function of [21].The partial likelihood function of the observed count time series Y_t is mathematically expressed as:

$$PL(\mu_t, \theta_t) = \sum_{i=1}^{N} \log\left[f_{Y_t}(y_t \mid \mathcal{F}_{t-1})\right] \tag{20}$$

The GARCH (1,1) parameters were estimated through the MLE technique on the standardized residuals extracted from the optimal count modeling procedure between the ZIPA and ZINBA models. This technique assumes that the model innovations follow a Gaussian distribution and estimates the model parameters by optimizing the likelihood or log-likelihood function associated with the assumed distribution. The estimation process was implemented using the $rugarch$ package in R, which offers a comprehensive and flexible environment for specifying and fitting GARCH models.

3.6. Model Selection

The model comparison criterion by [22] was utilised to compare ZIPA and ZINBA models. This is the most used technique for assessing whether ZIPA fits the zero-inflated overdispersed count time series than the ZINBA model [23]. The Vuong test statistic is expressed as follows:

$$Z = \frac{LR_N}{\sqrt{N}\,\omega_N} \tag{21}$$

where LR_N is the log-likelihood ratio between the ZINBA and ZIPA models, ω_N is the normalization factor, and N is the number of observations.

3.7. Model Evaluation Metrics

The study utilises the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) forecasting accuracy measurement functions to assess the performance of the fitted models. The formulas for computing the MAE and RMSE are expressed as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\varepsilon}_i|\,, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i^2} \tag{22}$$

The hat-epsilon $\hat{\varepsilon}_i$ denotes the difference between the actual and forecasted values, and it is mathematically expressed as:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \tag{23}$$

where Y_i and $\hat{Y}_i$ are the actual observed and forecasted values, respectively. The model with the smallest MAE and RMSE is generally accepted as the best performing model in terms of forecasting accuracy.

4. Results and Discussion

This section presents and discusses the empirical results of the data analysis conducted using the count time series models outlined in Section 3.

Table 4. Parameter estimates and their corresponding $p - values$ for the ZIPA model

Variable	**National**		**Kwazulu-Natal**		**Gauteng**		**Western Cape**	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Count model coefficients (Poisson model with log link)								
Intercept	4.1750	<0.0001	2.6500	<0.0001	3.2060	<0.0001	2.7119	<0.0001
Lag of 1	0.0049	<0.0001	0.0204	<0.0001	0.0129	<0.0001	0.0189	<0.0001
Trend	-0.0005	<0.0001	-0.0004	<0.0001	-0.0010	<0.0001	-0.0003	<0.0001
Zero-inflation model coefficients (binomial model with logit link)								
Intercept	-2.0876	0.0639	-0.8660	<0.0001	-0.3670	0.0639	-1.8471	<0.0001
Lag of 1	-0.0479	<0.0001	-0.1146	<0.0001	-0.0616	<0.0001	-0.0579	<0.0001
Trend	-0.0010	0.2439	0.0007	0.0499	-0.0019	<0.0001	0.0012	0.0098

Table 5. Parameter estimates and their corresponding $p - values$ for the ZINBA model

Variable	**National**		**Kwazulu-Natal**		**Gauteng**		**Western Cape**	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Count model coefficients (Negative Binomial model with log link)								
Intercept	3.8512	<0.0001	2.3832	<0.0001	3.0626	<0.0001	2.4237	<0.0001
Lag of 1	0.0071	<0.0001	0.0310	<0.0001	0.0262	<0.0001	0.0264	<0.0001
Trend	-0.0005	<0.0001	-0.0006	0.0120	-0.0019	<0.0001	-0.0004	0.0159
Log of Dispersion	0.3940	<0.0001	-0.3563	<0.0001	0.0744	0.1940	0.0949	0.1423
Zero-inflation model coefficients (Binomial model with logit link)								
Intercept	-1.1966	0.1200	-1.0997	0.0045	1.7134	<0.0001	-0.9924	0.0306
Lag of 1	-0.0414	0.5420	-0.3724	0.0003	-0.0029	0.7890	-0.3182	0.0100
Trend	-0.0299	0.2439	0.0005	0.4417	-0.0301	<0.0001	0.0001	0.8647

The ZIPA and ZINBA models in Table 4 and 5 were fitted using the lag of 1 and trend as predictor variables. These predictor variables play a very important role in explaining both the count and the zero-inflation processes. The lag of 1 captures the influence of the COVID-19 death count of the past day on the count of the current day, while the trend reflects the direction of the change in the COVID-19 death count over time.

In both the log-linear and logit components of the ZIPA model, the lag of 1 has the p-value less than the 5% level of significant across all the datasets. These results suggest that the log of the expected COVID-19 death count of today is influenced by the death count from the previous day. Furthermore, the probability of excess zeros in the COVID-19 death count of the current day is affected by proportion of excess zeros from the yesterday. It worth noting that in the log-linear component, the lag of 1 has a positive association with the log count of COVID-19 death cases, while in the zero-inflation (logit) component it has a negative association.

In the Poisson model with log link, it is found that the trend predictor significantly decreases the COVID-19 death log counts across all datasets. However, in the logit component of the ZIPA model the effect of the trend varies. That is, it has the decreasing insignificant effect at national level, increasing significant effect at KwaZulu Natal and Western Cape, and a decreasing significant effect in the Gauteng province. The intercept of the count model with a log link is positive and statistically significant across all datasets. In the logit component, it is significantly negative in KwaZulu-Natal and Western Cape, while negative but not statistically significant at the national level and in Gauteng province.

The predictor variables in the log-linear component of the ZINBA model depicted in Table 5 are statistically significant at 5% level across the four datasets. Trend predictor variable has a decreasing effect in the expected log of the COVID-19 death counts across all the levels (national and provincial). The logarithm of the dispersion parameter is statistically significant in the national level and KwaZulu Natal, while insignificant in Gauteng and Western Cape provinces. The natural logarithm of the dispersion parameter for the KwaZulu Natal dataset is negative which suggests that the underlying dispersion is less than 1. The significance of the parameter estimates in the zero-inflation component of the ZINBA model varies across datasets. In the national dataset, none of the estimates are significant at the 5% level. In KwaZulu-Natal and Western Cape, all parameter estimates are significant at the 5% level, except for the coefficient of the trend variable. Meanwhile, in Gauteng, the coefficient of the lag of 1 predictor is the only one which remained insignificant.

From a public health perspective, the findings in Tables 4 and 5 highlight key epidemiological patterns. The positive and statistically significant Lag of 1 coefficient in the log-linear components (count models) suggests strong temporal dependence, meaning yesterday's death count is a significant predictor of today's death count, reflecting the contagious nature and progression of the disease. Conversely, the negative Lag of 1 coefficient in the zero-inflation model suggests that higher death counts on the preceding day are associated with a lower likelihood of observing excess zeros, which may reflect improved consistency in reporting during periods of heightened transmission.

The consistently negative coefficient of the trend variable across all models indicates an overall downward trajectory in daily death counts, even in the presence of short-term spikes or pandemic waves. This sustained decline may reflect the cumulative impact of government interventions, such as lockdowns, widespread testing, and vaccination campaigns, and as well as shifts in public behavior, including increased social distancing, mask usage, and self-isolation practices.

Table 6. Vuong Test for Comparing Competing Models

Vuong Non-Nested Hypothesis Test-Statistic: (test-statistic is asymptotically distributed $N(0, 1)$ under the null that the models are indistinguishable)

Region	**Z-statistic**	**Alternative Hypothesis**	**P-value**
National	13.26851	ZINBA > ZIPA	<0.0001
Kwazulu-Natal	8.80473	ZINBA > ZIPA	<0.0001
Gauteng	9.345403	ZINBA > ZIPA	<0.0001
Western Cape	7.213288	ZINBA > ZIPA	<0.0001

The $p - value$ of the Vuong test in Table 6 is less than 0.01% across all the four datasets. These $p - values$ are all below the 5% level of significance, which suggests that the null hypothesis stating that the ZIPA and ZINBA models are indistinguishable should be rejected in favour of the alternative hypothesis that the ZINBA model is better than the ZIPA model. Therefore, it can be concluded with 95% confidence level that the ZINBA model performs better than the ZIPA model in the dataset with the characteristics similar to the ones observed in the COVID-19 death counts of South Africa and its three COVID-19 epicenter provinces.

The time series plots in Figure 2 exhibited clear signs of heteroscedasticity, indicating that the variability in COVID-19 death counts changed over time. The extracted residuals of the favored ZINBA model were found to exhibit ARCH effects, suggesting the presence of time-dependent volatility not captured by the count model alone. To address this, a two-step hybrid modeling approach was adopted: first, the ZINBA model was fitted to capture the structural features of the count data, and then a standard GARCH (1,1) model was applied to the residuals to model the conditional heteroscedasticity. This sequential strategy is pragmatic and computationally feasible, especially when dealing with complex count data. However, it is important to acknowledge that a single-stage joint estimation where both the ZINBA and GARCH components are estimated simultaneously would be statistically more efficient, as it allows for full likelihood-based inference and better integration of uncertainty across components. Despite its computational complexity, such an approach could offer deeper insights and more robust parameter estimates and may be worth exploring in future research.

Table 7. Test for ARCH Effects in the Residuals of the ZINBA Model

Null Hypothesis: No ARCH Effects								
Test	**National**		**Kwazulu-Natal**		**Gauteng**		**Western Cape**	
	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Lagrange Multiplier	33.718	0.0007	34.654	0.0005	63.398	<0.0001	109.6	<0.0001

The ZINBA model performed better than the ZIPA model across all the four datasets. Consequently, the residuals of the fitted ZINBA models were extracted and assessed for the existence of ARCH effects. The results of the ARCH-LM test presented in Table 7 reject the null hypothesis that there are no ARCH effects in the residuals. Therefore, it can be concluded with 95% confidence level that there are ARCH effects in the residuals of all the ZINBA models and they (residuals) can be utilised to fit the GARCH model.

Table 8. Parameter Estimates and P-values of the GARCH(1,1) Model

Parameter	**National**		**Kwazulu-Natal**		**Gauteng**		**Western Cape**	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
ω	0.0321	0.0098	0.0139	0.0083	0.1798	0.0019	0.3248	<0.0001
α_1	0.0454	0.0078	0.0910	<0.0001	0.1666	<0.0001	0.1540	<0.0001
β_1	0.9253	<0.0001	0.9079	<0.0001	0.8324	<0.0001	0.8607	<0.0001

Table 8 depicts the maximum likelihood estimates and corresponding $p-values$ of the GARCH (1,1) model fitted using the residuals of the ZINBA model. The $p-values$ of all the parameter estimates are less than the 5% level of significance, which suggests that the GARCH (1,1) model has statistically significant estimates. It is worth noting that the GARCH(1,1) model exhibits stationarity across all regions analyzed. This is confirmed by the condition $\alpha_1 + \beta_1 < 1$, which holds true for National $(0.0454 + 0.9253 = 0.9707)$, Kwazulu-Natal $(0.0910 + 0.9079 = 0.9989)$, Gauteng $(0.1666 + 0.8324 = 0.9990)$, and Western Cape $(0.1540 + 0.8607 = 1.0147)$. While Western Cape slightly exceeds the threshold, the model may still be considered weakly stationary depending on the context and estimation precision.

The sign-bias test by [20] was conducted to check whether the standard GARCH is well suited for the data or asymmetric GARCH-type model should be considered. The results (insignificant p-values) depicted in Table 9 fail to reject the null hypothesis that there are no asymmetric effects in the data. Therefore, these findings support the adequacy of the fitted standard GARCH model presented in Table 8.

Table 9. Sign-Bias Test for Asymmetric Effects

Null Hypothesis: No asymmetric effects		
Region	**Statistic**	**P-value**
National	1.7740	0.0765
Kwazulu-Natal	0.1429	0.8864
Gauteng	1.0192	0.3084
Western Cape	0.3900	0.6967

Table 10. Performance Metrics of ZINBA and ZINBA_GARCH Models

Model	**National**		**Kwazulu-Natal**		**Gauteng**		**Western Cape**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ZINBA	77.3673	173.7849	16.2608	36.4017	18.0112	40.3180	18.8406	47.7018
ZINBA_GARCH	77.5891	173.9354	17.3431	36.6225	21.5529	41.6128	26.9054	51.2029

The forecasting accuracy of the fitted ZINBA and ZINBA-GARCH (1,1) models were measured by the MAE and RMSE depicted in Table 9. The proposition of these accuracy measures is parallel to the suggestions of the time series plots presented in Figure 2. That is, the ZINBA model is deemed as the most desirable forecasting approach across all the four considered datasets compared to incorporating volatility through the GARCH components. However, the Lagrange Multiplier (ARCH-LM) test performed on the residuals of the ZINBA model Table 7 warranted the incorporation of the GARCH components. Therefore, using the ZINBA-GARCH (1,1) makes sense despite the slightly higher error metrics, especially in the Western Cape Gauteng Province.

Figure 3 depicts the time series plots of the daily COVID-19 death counts, comparing observed data with forecasts at the national level and across three key epicenter provinces in South Africa. The black solid line represents the actual observed COVID-19 death counts, while the red and blue lines denote the forecasts generated counts using the fitted ZINBA and ZINBA-GARCH (1,1) models, respectively. Looking at the Figure, it is observed that the three time series are mostly consistent. However, it worth noting that at national level and in KwaZulu Natal, the ZINBA and ZINBA-GARCH (1,1) models behave similarly. Moreover, In Gauteng and Western Cape, the ZINBA model appears to perform better in certain instances as compared to the ZINBA-GARCH (1,1) model. appears to overfit as compared to the ZINBA model, the trends are similar. These results suggests that the GARCH components incorporated in the ZINBA model might not be adding any value in the Gauteng and Western Cape datasets.

The LagEffect plot (top-left with green color) reflects how the COVID-19 death counts of yesterday influence today's death counts. The sharp spikes suggest that certain periods had strong dependence on the previous day's death counts. These periods of sharp spikes are likely to be during the different waves of the COVID-19 pandemic in South Africa.

The trend plot (top-right with blue color) depicts the directional movement of the South African COVID-19 deaths cases, independent of the autocorrelation and zero inflation. The plot depicts a clear downward slope suggesting that the expected number of COVID-19 deaths is decreasing over time. This decline could be attributed to the increased public awareness, vaccination programmes, and the transition from pandemic to endemic.

The volatility plot (bottom-left with purple color) reflects the conditional variance from the GARCH (1,1) model fitted on the residuals extracted from the ZINBA model. The fluctuations depicts that the uncertainty in the COVID-19 death counts in South Africa varies over time. The period of high volatility reflects the unpredictability of the pandemic's impact on the South Africa's COVID-19 mortality.

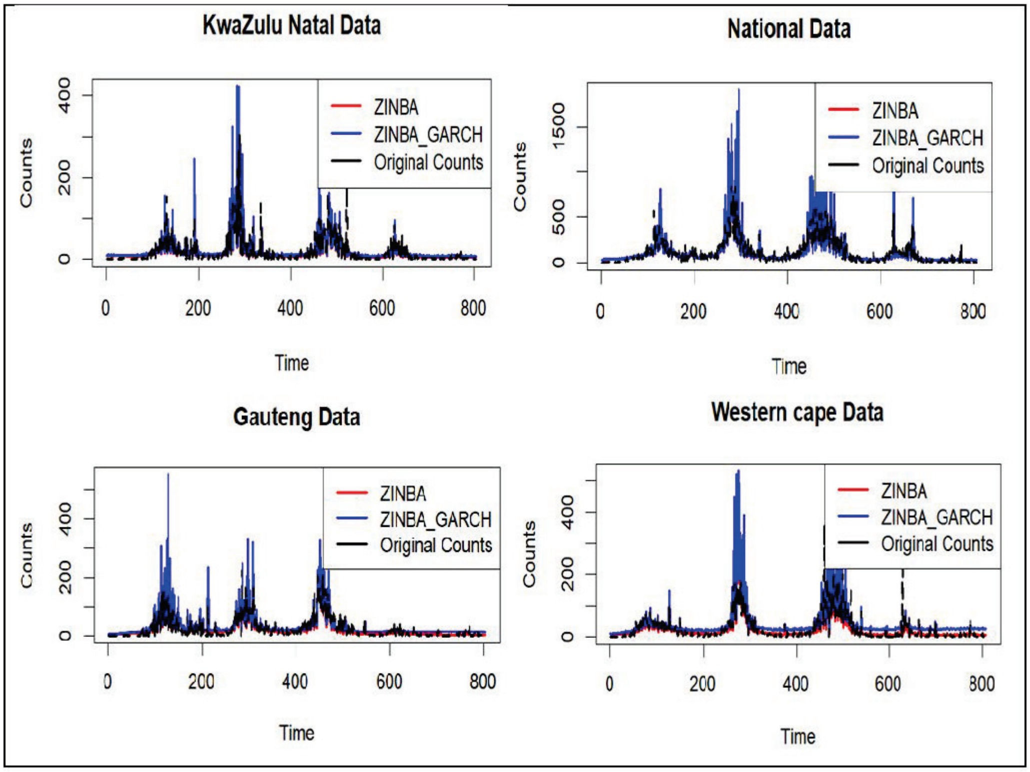

Figure 3. Plots for the ZINBA, ZINBA-GARCH (1,1) forecasts and the original death counts.

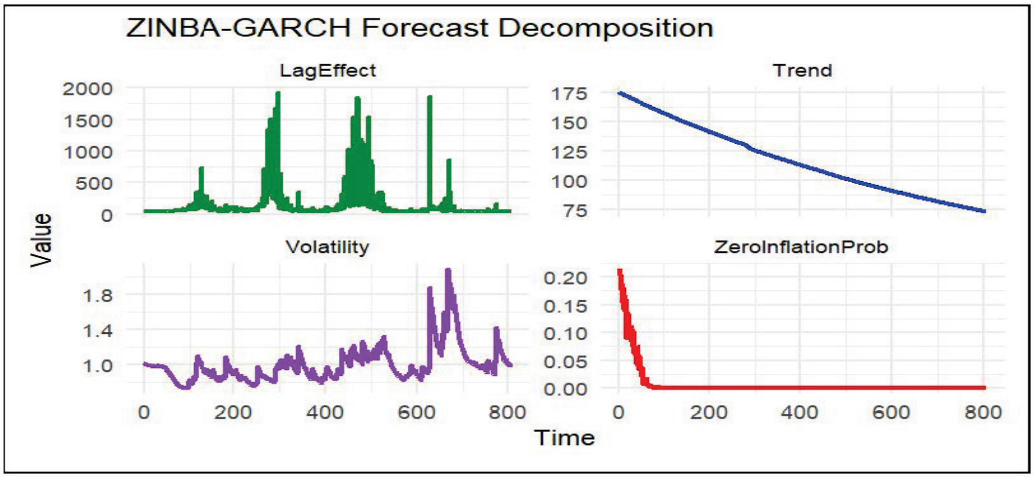

Figure 4. Forecast decomposition of COVID-19 death counts in South Africa using the ZINBA-GARCH framework.

The zero-inflation probability plot (bottom right in red color) illustrate the probability that a COVID-19 death counts in South Africa is a structural zero not just a random low value. The high initial probability dropping to near zero suggests that early in the pandemic, many days had no reported COVID-19 death counts, possibly due to limited testing, underreporting, or the virus not yet spreading widely.

5. Conclusion

The South African COVID-19 datasets exhibited excess zeros, overdispersion, autocorrelation, and heteroscedasticity. This study utilised ZIPA and ZINBA models to address the overdispersed count data with excessive zeros. The Vuong test favored the ZINBA model over ZIPA, and the sign-bias test results aligned with Mthethwa et al. (2022), indicating that the standard GARCH (1,1) specification is appropriate for capturing volatility in these datasets. Consequently, residuals from the ZINBA model were used to fit the GARCH (1,1) component.

Interestingly, while the ZINBA-GARCH (1,1) model is statistically more comprehensive, accounting for conditional heteroscedasticity, the forecasting accuracy metrics (MAE and RMSE) consistently favored the simpler ZINBA model. This reveals a critical tension between model adequacy and predictive performance. The presence of ARCH effects in the ZINBA residuals justifies the inclusion of GARCH dynamics; however, it is possible that the GARCH component captures in-sample volatility patterns that do not translate into improved out-of-sample forecasts. Alternatively, the added complexity may introduce overfitting, particularly in provinces like Western Cape and Gauteng where volatility is more pronounced.

Therefore, while ZINBA-GARCH (1,1) may be the statistically correct model for representing the data-generating process, the ZINBA model may be sufficient and even preferable for pure point forecasting. This distinction is crucial for practitioners: model selection should be guided not only by statistical diagnostics but also by the intended application. A more nuanced conclusion is that both models offer value, depending on whether the goal is structural understanding or forecasting precision.

Declaration

Ethics statement: Ethical clearance is not required as the data is publicly accessible online.

Data availability: The data is available for access on the website: https://github.com/dsfsi/covid19za/blob/master/data/

Acknowledgement

The authors gratefully acknowledge the insightful and constructive comments provided by the anonymous reviewers, and the Editor, which greatly contributed to improving the clarity and overall quality of this manuscript.

REFERENCES

1. F. A. Chyon, M. N. H. Suman, M. R. I. Fahim, and M. S. Ahmmed, *Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning*, Journal of Virological Methods, vol. 301, p. 114433, 2022.
2. S. S. Januri, I. A. Malek, N. Nasir, and Z. A. M. M. Yasin, *Forecasting the Spread of Daily Confirmed COVID-19 Cases in Malaysia*, International Journal of Academic Research in Business and Social Sciences, vol. 12, no. 2, pp. 322–334, 2022.
3. F. Petropoulos, S. Makridakis, and N. Stylianou, *COVID-19: Forecasting confirmed cases and deaths with a simple time series model*, International Journal of Forecasting, vol. 38, no. 2, pp. 439–452, 2022.
4. V. Vig and A. Kaur, *Time series forecasting and mathematical modeling of COVID-19 pandemic in India: a developing country struggling to cope up*, International Journal of System Assurance Engineering and Management, vol. 13, no. 6, pp. 2920–2933, 2022.
5. S. M. Alzahrani, *A log linear Poisson autoregressive model to understand COVID-19 dynamics in Saudi Arabia*, Beni-Suef University Journal of Basic and Applied Sciences, vol. 11, no. 1, p. 118, 2022.
6. N. Mthethwa, R. Chifurira, and K. Chinhamu, *Estimating the risk of SARS-CoV-2 deaths using a Markov switching-volatility model combined with heavy-tailed distributions for South Africa*, BMC Public Health, vol. 22, no. 1, p. 1873, 2022.
7. I. P. Ratnayake and V. A. Samaranayake, *An integer GARCH model for a Poisson process with time-varying zero-inflation*, PLOS ONE, vol. 18, no. 5, p. e0285769, 2023.

8. 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, pp. 1–24, 2023.
9. K. Tawiah, W. A. Iddrisu, and K. Asampana Asosega, *Zero-Inflated Time Series Modelling of COVID-19 Deaths in Ghana*, Journal of Environmental and Public Health, vol. 2021, no. 1, p. 5543977, 2021.
10. M. Alexander, *The population of South Africa's nine provinces*, 2024. Available at: https://images.app.goo.gl/QHDkgu7XsfWa9psx6.
11. K. K. Yau, A. H. Lee, and P. J. Carrivick, *Modeling zero-inflated count series with application to occupational health*, Computer Methods and Programs in Biomedicine, vol. 74, no. 1, pp. 47–52, 2004.
12. J. H. Jakobsen, *Application of count time series to battle deaths*, Master's thesis, University of Oslo, 2021.
13. M. Yang, *Statistical models for count time series with excess zeros*, Doctoral dissertation, The University of Iowa, 2012.
14. 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, vol. 14, pp. 26–38, 2013.
15. D. Lambert, *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, vol. 34, no. 1, pp. 1–14, 1992.
16. A. H. Lee, K. Wang, and K. K. Yau, *Analysis of zero-inflated Poisson data incorporating extent of exposure*, Biometrical Journal, vol. 43, no. 8, pp. 963–975, 2001.
17. T. Loeys, B. Moerkerke, O. De Smet, and A. Buysse, *The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression*, British Journal of Mathematical and Statistical Psychology, vol. 65, no. 1, pp. 163–180, 2012.
18. T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, Journal of Econometrics, vol. 31, no. 3, pp. 307–327, 1986.
19. R. F. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*, Econometrica: Journal of the Econometric Society, pp. 987–1007, 1982.
20. R. F. Engle and V. K. Ng, *Measuring and testing the impact of news on volatility*, The Journal of Finance, vol. 48, no. 5, pp. 1749–1778, 1993.
21. B. Kedem and K. Fokianos, *Regression models for time series analysis*, John Wiley & Sons, 2005.
22. Q. H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica: Journal of the Econometric Society, pp. 307–333, 1989.
23. B. A. Desmarais and J. J. Harden, *Testing for zero inflation in count models: Bias correction for the Vuong test*, The Stata Journal, vol. 13, no. 4, pp. 810–835, 2013.