Scientific Research Publishing

Open Journal of Modelling and Simulation, 2023, 11, 14-36
https://www.scirp.org/journal/ojmsi
ISSN Online: 2327-4026
ISSN Print: 2327-4018

Modelling COVID-19 Cumulative Number of Cases in Kenya Using a Negative Binomial INAR (1) Model

Charity Wamwea, Susan Mwelu, Matabel Odin

Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya
Email: wamweac@gmail.com

How to cite this paper: Wamwea, C., Mwelu, S. and Odin, M. (2023) Modelling COVID-19 Cumulative Number of Cases in Kenya Using a Negative Binomial INAR (1) Model. *Open Journal of Modelling and Simulation*, **11**, 14-36.
https://doi.org/10.4236/ojmsi.2023.111002

Received: August 26, 2022
Accepted: January 27, 2023
Published: January 30, 2023

Abstract

In this paper, a Negative Binomial (NB) Integer-valued Autoregressive model of order 1, INAR (1), is used to model and forecast the cumulative number of confirmed COVID-19 infected cases in Kenya independently for the three waves starting from 14th March 2020 to 1st February 2021. The first wave was experienced from 14th March 2020 to 15th September 2020, the second wave from around 15th September 2020 to 1st February 2021 and the third wave was experienced from 1st February 2021 to 3rd June 2021. 5, 10, and 15-day-ahead forecasts are obtained for these three waves and the performance of the NB-INAR (1) model analysed.

Keywords

COVID-19 Predictive Model, New SARS-CoV-2, Integer Valued Autoregressive (INAR) Model

1. Introduction

Coronaviruses (CoV) are a large family of zoonotic viruses that are transmitted from animal to human and cause symptoms ranging from sneezing, fever, headaches, pneumonia and severe breathing difficulties. The new strain of coronaviruses, SARS-CoV-2, better known as COVID-19 was first identified in humans in December 2019 in Wuhan, China. The first COVID-19 positive case in Kenya was confirmed on 14th March 2020. The rise in positive cases resulted in enforcement of stringent measures such as closure of all social amenities, banning of both domestic and international flights, lockdown of the Nairobi metropolitan area and Mombasa and a nationwide night curfew (1900 hrs - 0500 hrs) by the government of Kenya.

Though most infected cases are either asymptomatic or could be treated at home with the use of over the counter medicines, a few turn out to be tragic leading to death. Therefore, in order to protect oneself and others from COVID-19 infection, the world health organization (WHO) recommended that the general public should:

- Regularly and thoroughly wash their hands with soap and water, or use an alcohol-based hand sanitizer.
- Maintain a distance of at least 1 meter (5 feet) from each other.
- Stay at home or keep a social distance if one has COVID-19 symptoms such as coughing or sneezing and avoid mixing with others in a crowd.
- Maintain good respiratory hygiene by covering the mouth and nose while coughing and sneezing with a handkerchief, tissue, or into a flexed elbow.
- Suspend all public gatherings, meetings, religious crusades games events etc.
- Attend religious services only if they provide sanitizing or hand washing.
- Suspend all inter-school events, but keep schools open.
- Ensure public transport providers provide hand sanitizers for their clients and regular cleaning of their vehicles.
- Temporarily suspend prison visits.
- Restrict oneself from traveling outside the country unless when it is absolutely necessary and no travel to disease epicenter countries.

With the visible negative impacts these measures had on the Kenyan economy, the GoK had to reduce some of these stringent measures. For instance, on 6th July 2020, the lockdown on the two major cities of Mombasa and Nairobi was removed. In addition to this, the curfew hours were reduced such that they would be in effect every day from 2300 hrs to 0400 hrs. As at the time, Kenya had a total of 7886 confirmed COVID-19 positive cases and 160 deaths respectively. The international ban on flights was later on lifted in August 2020 and as of 11th November 2020, having no known vaccine for the SARS-CoV-2, 1,285,013 individuals had lost their lives due to COVID-19 worldwide. In Kenya, the total number of confirmed positive cases and deaths as of the same date was 64,588 and 1154 respectively. It was now evident that a second wave of the infection had hit all the areas of the world due to the relaxation of the stringent measures. This led to the need for researchers to come up with predictive mathematical models.

Mathematical modeling and prediction of positive COVID-19 cases helps governments to know the expected severity of the disease in advance allowing them to promptly respond, prepare and give guidelines towards the mitigation and reduction of infections. Several COVID-19 models have been proposed to model different COVID-19 aspects so far. For example, [1] used different ARIMA models to estimate the COVID-19 prevalence in Italy, Spain and France for the period between 21st February 2020 and 15th April 2020. His fitted models had a mean absolute percentage error (MAPE) values ranging from 4.75% to 5.63%.

[2] used 3 different models: Grey (1, 1) model, Non-linear Grey Bernoulli (1, 1) model and the Fractional Non-Linear Grey Bernoulli (1, 1) model to forecast the

cumulative number of confirmed COVID-19 cases in Italy, UK and USA. They trained their model using data spanning from 19th March 2020 to 22nd April 2020 and tested their model using remaining data spanning from 23rd April 2020 to 22nd May 2020. Their MAPE values performed for the 1 month ahead forecasts for the three different countries were found to be 0.9%, 2.8% and 4.9% respectively.

[3] proposed an adaptive neuro-fuzzy inference system (ANFIS) using an enhanced flower pollination algorithm (FPA) by using a salp swarm algorithm (SSA) model to get a 7-day-ahead forecast of the number of confirmed COVID-19 cases during the first phase of the pandemic from 11th February 2020 to 18th February 2020. They compared the forecasting power of their model FPASSA with several other models: ANN, KNN, SVR, ANFIS, PSO, GA, ABC, and FPA and found theirs to be the best model with a MAPE of 4.8%.

[4] used an ARIMA model to estimate and forecast the number of confirmed COVID-19 cases for 5 countries: India, Brazil, Russia, Spain and USA. The MAPE values for the different countries were calculated from the 18-day-ahead forecasts from 1st July 2020 to 18th July 2020 and their values were found to be 3.7%, 1.8%, 1.1%, 0.8% and 2.9% respectively.

[5] used different time-series and curve estimation regression models to forecast the number of COVID-19 new cases in 10 African countries: South Africa, Egypt, Morocco, Ethiopia, Nigeria, Algeria, Ghana, Kenya, Cameroon and Cote-divore. His forecast period was from 14th February 2020 to 6th September 2020. The MAPE values for the different fitted models ranged from 14.6% to 191.19% with the best Kenyan model having an in-sample MAPE of 57.39%.

[6] forecasted the number of confirmed COVID-19 cases in Italy, Spain, France, China, Australia and USA using several models: RNN, GRU, LSTM, BiLSTM and VAE. The MAPE values of the different models were calculated from the 17 day-ahead forecasts from 1st June 2020 to 17th June 2020. The best model out of the five was found to be the VAE. The VAE MAPE values for the different countries were found to be 5.9%, 2.2%, 1.9%, 0.1%, 0.2% and 2.0% respectively.

[7] used different ARIMA models to project the COVID 19 prevalence pattens in Ethiopia, Djibouti, Sudan and Somalia. The dataset considered for the study was from 13th March 2020 to 30th June 2020. The MAPE values of the different fitted models ranged from 3.59% to 3.92%.

[8] used a log-polynomial model to forecast the ratio of the number of daily new diagnosed cases combined with an INAR (1) model to forecast the number of confirmed new cases of COVID-19 in Italy for the period between 19th May 2020 to 2nd June 2020. Their results were then compared with those of the ARIMA model. They found that their proposed model outperformed the ARIMA model. The mean absolute error (MAE) values of the fitted model ranged from 47.56 to 53.05 for the different h-days forecasts.

In this paper, a first-order mathematical integer-valued autoregressive INAR (1) model is used to predict and forecast the evolution of the total number of

positive confirmed COVID-19 cases in Kenya across the first three waves for the period between 1st May 2020 to 31st March 2021. The INAR model first proposed by [9]; [10] is a Markov model used to model stationary count processes with discrete marginal distributions. Count data expresses the number of certain units or events in a specified context. Its possible outcomes are contained in the set of non-negative integers, $\mathbb{N}_0 = \{0,1,2,\cdots\}$. The INAR model has extensive practical applications in different fields of science. Daily COVID-19 Kenyan data freely collected from the website: https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data was used in this research.

2. Methods

2.1. The Model

Let X be a discrete count random variable with a range $\mathbb{N}_0$ and $\alpha \in (0,1)$ be a random variable such that αoX is given by:

$$\alpha oX = \sum_{i=1}^{X} Z_i \tag{1}$$

arises from X by binomial thinning and Z_i's are independent and identically distributed (iid) binary random variables with

$$\alpha = Pr(Z_i = 1) \tag{2}$$

which are independent of X.

Definition: First-Order Integer-Valued Autoregressive INAR (1) Model

Let $(\varepsilon_t)_{\mathbb{N}_0}$ be an iid process with a range $\mathbb{N}_0$ such that

$$E[\varepsilon_t] = \mu_\varepsilon \quad \text{and} \quad Var(\varepsilon_t) = \sigma_\varepsilon^2 \tag{3}$$

and $\alpha \in (0,1)$, then a process $(X_t)_{\mathbb{N}_0}$ of observations that follow the recursion:

$$X_t = \alpha oX_{t-1} + \varepsilon_t \tag{4}$$

is said to be a first-order integer valued autoregressive INAR (1) process if all thinning operations are performed independently of each other and of $(\varepsilon_t)_{\mathbb{N}_0}$.

Remark. It should be noted that the possible outcomes of both the daily number of confirmed COVID-19 cases and the cumulative number of confirmed COVID-19 cases contain non-negative integer values, $\mathbb{N}_0$. Therefore, to use the INAR (1) to model the total number of confirmed positive COVID-19 cases, the following assumptions were made:

X_t: Total population of people at time, t, that have tested positive for COVID-19.

X_{t-1}: Total population of people at time, $t-1$, that have tested positive for COVID-19.

ε_t: Population of people at time, t, that have tested positive for COVID-19.

The conditional mean and variance of X_t given X_{t-1} are:

$$E\left[X_t \mid X_{t-1}\right] = \alpha o X_{t-1} + \mu_\varepsilon \tag{5}$$

$$Var\left(X_t \mid X_{t-1}\right) = \alpha\left(1-\alpha\right)X_{t-1} + \sigma_\varepsilon^2 \tag{6}$$

Assuming that

$$E\left[X_t\right] = \mu \quad \text{and} \quad Var\left(X_t\right) = \sigma^2 \tag{7}$$

then;

$$E\left[\alpha o X_t\right] = \alpha\mu \tag{8}$$

$$Var\left(\alpha o X_t\right) = \alpha^2\sigma^2 + \alpha\left(1-\alpha\right)\mu \tag{9}$$

The unconditional mean and variance of X_t are:

$$E\left[X_t\right] = E\left[\alpha o X_{t-1} + \varepsilon_t\right] = \mu = \alpha\mu + \mu_\varepsilon \tag{10}$$

Hence

$$\mu = \frac{\mu_\varepsilon}{1-\alpha} \tag{11}$$

and

$$Var\left(X_t\right) = Var\left[\alpha o X_{t-1} + \varepsilon_t\right] = \sigma^2 = \alpha^2\sigma^2 + \alpha\left(1-\alpha\right)\mu + \sigma_\varepsilon^2 \tag{12}$$

Hence

$$\sigma^2 = \frac{\sigma_e^2 + \alpha\mu_\varepsilon}{1-\alpha^2} \tag{13}$$

The autocorrelation function (ACF) of a stationary INAR (1) process is given by:

$$\rho(k) = corr\left(X_t, X_{t-k}\right) = \alpha^k \tag{14}$$

The 1-step transitional probabilities of an INAR (1) model is given by:

$$p_{k|l} = Pr\left(X_t = k \mid X_{t-1} = l\right) \tag{15}$$

$$= \sum_{j=0}^{\min(k,l)} \binom{l}{j} \alpha^j \left(1-\alpha\right)^{l-j} \cdot P\left(\varepsilon_t = k - j\right) \tag{16}$$

The index of dispersion of the innovations ε_t is given by:

$$I_\varepsilon = \frac{\sigma_\varepsilon^2}{\mu_\varepsilon^2} \in (0, \infty) \tag{17}$$

such that if:

$$\begin{cases} I_\varepsilon = 1, & \text{the variable is equidispersed} \\ I_\varepsilon < 1, & \text{the variable is underdispersed} \\ I_\varepsilon > 1, & \text{the variable is overdispersed} \end{cases}$$

The index of dispersion of X_t is given by:

$$I = \frac{\sigma^2}{\mu} = \frac{I_\varepsilon + \alpha}{1+\alpha} \tag{18}$$

Hence

$$I_e = I(1+\alpha) - \alpha \tag{19}$$

Remark. The innovations are usually used to determine the type of INAR model to be fitted.

2.1.1. Poisson INAR (1) Model

A Poisson (α, λ) INAR model is fitted to the data if the innovations, ε_t, are Poisson (λ) distributed. Poisson distributed random variables are usually equi-dispersed hence $I = 1$ if $\varepsilon_t \sim \text{Poisson}(\lambda)$. Under the Poisson model, the mean and variance of the innovations is given by:

$$\mu_\varepsilon = \sigma_\varepsilon^2 = \lambda \tag{20}$$

2.1.2. Negative Binomial INAR (1) Model

A negative binomial (α, r, p) INAR model is fitted to the data if the innovations, ε_t, are negative binomially NB (r, p) distributed. Negative binomially distributed random variables are usually overdispersed, hence $I > 1$. Under the negative binomial model, the mean and variance of the innovations are given by:

$$\mu_\varepsilon = \frac{rp}{q} \quad \text{and} \quad \sigma_\varepsilon^2 = \frac{rp}{q^2} \tag{21}$$

where $q = 1 - p$

2.1.3. Geometric INAR (1) Model

A geometric (α, p) INAR model is fitted to the data if the innovations, ε_t, are geometrically (p) distributed,. Geometrically distributed random variables are usually underdispersed, hence $I < 1$. Under the geometric model, the mean and variance of the innovations are given by:

$$\mu_\varepsilon = \frac{q}{p} \quad \text{and} \quad \sigma_\varepsilon^2 = \frac{q}{p^2} \tag{22}$$

2.2. Model Identification

To identify the appropriate INAR (p) model, the following procedures must be undertaken.

2.2.1. ACF Structure

The ACF structure of the data is first analysed through the use of the autocorrelation (ACF) and partial autocorrelation (PACF) plots. Essentially, autocorrelation measures the relationship between a variable's current value and its past values. The sample autocorrelations of order k are calculated using the formula

$$\rho(k) = \frac{\frac{1}{n-k}\sum_{t=k+1}^{n}\left(X_t - \bar{X}\right)\left(X_{t-k} - \bar{X}\right)}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \bar{X}\right)^2}\sqrt{\frac{1}{n-k}\sum_{t=k+1}^{n-k}\left(X_{t-k} - \bar{X}\right)^2}}; \quad k = \{0,1,2,\cdots\} \tag{23}$$

where

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t \tag{24}$$

Partial autocorrelations measures the linear dependence of one variable after removing the effect of other variable (s) that affect both variables. The partial autocorrelation (PACF) of order k is obtained as a series of regressions of the form:

$$\tilde{X}_t = \phi_{k1}\tilde{X}_{t-1} + \phi_{k2}\tilde{X}_{t-2} + \cdots + \phi_{kk}\tilde{X}_{t-k} + e_t; \quad k = \{0,1,2,\cdots\} \tag{25}$$

where ϕ_{ki} is the i^{th} PACF value for a data with k lags and

$$\tilde{X}_t = X_t - \bar{X} \tag{26}$$

The ACF and PACF plots are usually used to determine the moving average (MA) and autoregression (AR) orders respectively.

2.2.2. Test for Stationarity

The data is then tested for stationarity. The Augmented Dickey Fuller (ADF) test can be used to achieve this. The test hypotheses considered while performing the ADF test is usually given by:

H_0: Series is not stationary (There is a unit root) *i.e.* $\gamma = 0$;

H_1: Series is stationary (There is no unit root) *i.e.* $\gamma < 0$.

The ADF test statistic is given by:

$$\text{ADF} = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{27}$$

The null hypotheses is rejected when the p-value < α, where α is the level of significance.

2.3. Parameter Estimation

The parameters of the chosen INAR (1) model was estimated using the method of moments approach.

Method of Moments (MOM) Approach

Let $X_1, X_2, \cdots, X_T$ be a time series from a stationary INAR (1) process. Under the MOM method, the true moments are replaced by the corresponding sample moments. That is:

$$\hat{\mu}_{MM} = \bar{X} \tag{28}$$

$$\hat{\alpha}_{MM} = \hat{\rho}(1) \tag{29}$$

$$= \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} \tag{30}$$

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{t=k+1}^{n}\left(X_t - \bar{X}\right)\left(X_{t-k} - \bar{X}\right); \quad k \in \mathbb{N}_0 \tag{31}$$

where $\bar{X}$ is the sample mean of the data as defined by Equation (24).

- For a Poisson (α, λ) INAR (1) model;

$$\begin{aligned} \hat{\alpha}_{MM} &= \hat{\rho}(1) \\ \hat{\lambda}_{MM} &= \hat{\mu}_{\varepsilon} = \bar{X}\left(1 - \hat{\alpha}_{MM}\right) \end{aligned} \tag{32}$$

- For a Negative Binomial (α, r, p) INAR (1) model;

$$
\begin{aligned}
\hat{\alpha}_{MM} &= \hat{\rho}(1) \\
\hat{p}_{MM} &= 1-\hat{q}_{MM} \\
\hat{q}_{MM} &= \frac{1}{\hat{I}_{\varepsilon}} \\
\hat{I}_{\varepsilon} &= \hat{I}_{MM}\left(1+\hat{\alpha}_{MM}\right)-\hat{\alpha}_{MM} \\
\hat{I}_{MM} &= \frac{\hat{\sigma}^2_{MM}}{\hat{\mu}_{MM}} \\
\hat{\sigma}^2_{MM} &= \frac{1}{T}\sum_{t=1}^{T}\left(X_t-\bar{X}\right)^2 \\
\hat{r}_{MM} &= \frac{\hat{\mu}_{\varepsilon}\times\hat{q}_{MM}}{1-\hat{q}_{MM}} \\
\hat{\mu}_{\varepsilon} &= \bar{X}\left(1-\hat{\alpha}_{MM}\right)
\end{aligned}
\tag{33}
$$

2.4. Forecasting the INAR (1) Process

Given the INAR (1) process, one can forecast the future outcomes of the process X_{t+h} forecast for some lag $h \geq 1$ using the observations $\{X_1, X_2, \cdots, X_T\}$. For real valued processes, one can use the conditional mean as it can yield an optimal value. Therefore, by applying the law of total expectation iteratively, it follows that the h-step ahead conditional mean is given by:

$$
E\left[X_{t+h} \mid X_t\right] = \alpha^h X_t + \mu_e\left(\frac{1-\alpha^h}{1-\alpha}\right) \tag{34}
$$

However, the main disadvantage of using this method is that it may lead to non-integers.

2.5. Model Adequacy

Model adequacy checks whether the fitted model is really adequate for the analyzed data. That is, if the resultant time series constitutes the typical realization of the considered model. This can be achieved using several procedures.

2.5.1. ACF Structure

The ACF structures of the test sample and the out-of-sample forecasts of the data can be compared. This was achieved by plotting their ACF and PACF values.

2.5.2. Marginal Characteristics

A comparison of the marginal characteristics of the test sample data and the out-of-sample forecasts can be performed for the mean and cumulative probabilities. This can be done by comparing its mean and dispersion ratios. In addition to this, the actual and forecasted probability distributions can be compared using the two-sample Kolmogorov-Smirnov test.

Kolmogorov Smirnov (K-S) Goodness of Fit Test

The Kolmogorov Smirnov (K-S) goodness of fit test is a non-parametric test of equality that can be used to compare a sample with a reference probability distribution. The K-S test was first introduced by [11]. The test is based on the

maximum difference between an empirical and a hypothetical cumulative distribution. The test hypotheses under this test are:

H_0: The sample is drawn from the reference distribution;

H_1: The sample is not drawn from the reference distribution.

Assuming that the empirical distribution function $F_n(x)$ for n independently and identically distributed ordered observations X_i defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i) \tag{35}$$

where

$$I(X_i) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{elsewhere} \end{cases} \tag{36}$$

Then, the K-S test statistic for a given cumulative distribution function $F(x)$ is given by:

$$D = \sup_x \left| F_n(x) - F(x) \right| \tag{37}$$

where $\sup_x$ is the supremum of the set of distances.

The null hypothesis is rejected if the p-value < α or if $D > c(\alpha)\sqrt{\frac{2}{n}}$ where

$$c(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \times \frac{1}{2}} \tag{38}$$

where n is the sample size.

2.5.3. Normality Assumption

According to the central limit theorem, if the data is large enough, then all distributions tend towards the normal distribution. For a large dataset, the standardized Pearson residuals of the fitted model against the test sample given by

$$R_t = \frac{e_t - \bar{e}}{s_e} \tag{39}$$

can be analyzed to confirm that they are normally distributed with mean 0 and variance 1, where:

$\bar{e}$ is the sample mean residuals;

s_e is the sample residual standard deviation.

To test for normality, the Shapiro Wilk (S-W) test is used. The S-W test hypotheses considered are usually:

H_0: Data is normally distributed.

H_1: Data is not normally distributed.

The S-W test statistics is given by:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \tag{40}$$

where:

$X_{(i)}$: i^{th} order statistic.

$\overline{X}$: Sample mean.

$$a = \left(a_1, a_2, \cdots, a_n\right) = \frac{m^{\mathrm{T}} V^{-1}}{C}$$

V: Covariance matrix.

$$C = \left(m^{\mathrm{T}} V^{-1} V^{-1} m\right)^{\frac{1}{2}}$$

$$m = \left(m_1, m_2, \cdots\right)^{\mathrm{T}}$$

is the expected values of the order statistics.

The null hypotheses is rejected when the p-value < α, where α is the level of significance.

3. Main Results

The cumulative total confirmed positive COVID-19 cases have been on the rise since the first COVID-19 case was identified on 14th March 2020. By 31st March 2021, a total of 133,895 people had tested positive for COVID-19 in Kenya. Figure 1 below shows this evolution.

The steady increase in the total number of positive COVID-19 cases can be attributed to the increase in the number of cases confirmed daily. Such increases or decreases in the daily number of cases are directly attributable to the lax of stringent measures given by the ministry of health to the people of Kenya. Figure 2 below is a plot of the number of confirmed positive COVID-19 cases each day from 14th March 2020 to 31st March 2021.

From Figure 2, it can be seen that the first COVID-19 wave started from March 2020 and ended in mid-September 2020 where the curve was flattened. The second wave started in September 2020 to the first week of January 2021 where the curve was flattened for the second time. The third wave started from January 2021 to date (31st March 2021) where the curve seems to be at its peak. The Kenyan COVID-19 dataset was therefore analyzed independently over the three waves from 1st May 2020 through to 31st March 2021. Throughout this paper, a 5% significance level is assumed.

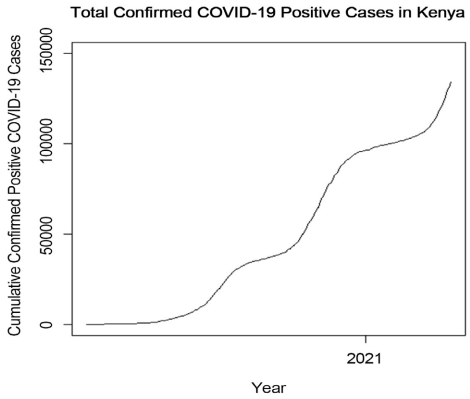


Figure 1. Total cumulative confirmed positive COVID-19 cases in kenya.

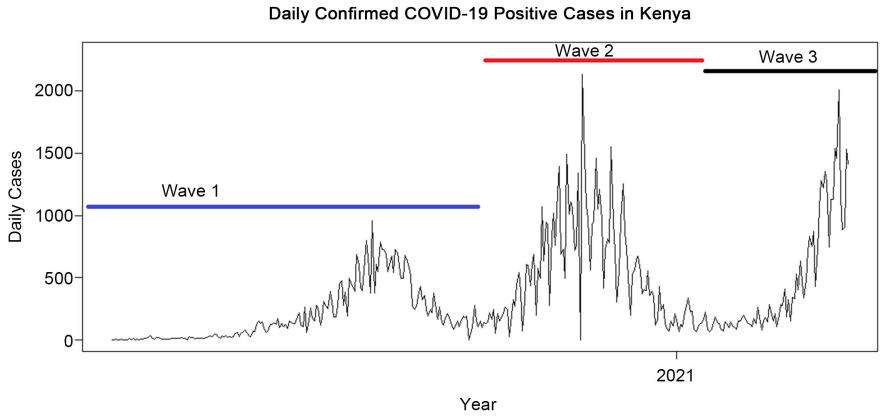


Figure 2. Daily confirmed positive COVID-19 cases in kenya.

3.1. Model Identification

The Kenyan data used to train the model was sub-divided and used to model the cumulative COVID-19 positive cases for the three waves experienced so far. Datasets for the period between 1st May 2020 to 11th September 2020, 22nd September 2020 to 23rd December 2020 and 3rd January 2021 to 1st March 2021 were used to train the first, second and third COVID-19 waves in Kenya. The appropriate INAR (p) model was considered for this study.

Autocovariance Structures The ACF and PACF plots for the three waves were plotted to help us identify the respective AR and MA orders. **Figure 3** below shows these plots.

From **Figures 3(a)-(c)** respectively, it can be seen that the ACF plots for the three waves decrease exponentially. However, there is one significant lag for the PACF plots. The cumulative COVID-19 positive confirmed cases data is an integer valued time-series dataset and hence with the help of **Figure 3**, we concluded that an INAR (1) model is appropriate to model the Kenyan scenario.

3.2. Descriptive Statistics

The descriptive Statistics of the train Kenyan data was as summarized in **Table 1** below.

From the descriptive statistics given in **Table 1**, it can be seen that for all the three waves, the mean of the data is less than its variance. This meant that the cumulative COVID-19 positive confirmed cases in Kenya are over-dispersed. Hence, the choice of the negative binomial INAR (1) model was used to model the Kenyan COVID-19 data in this paper.

3.3. Testing for Stationarity

The ADF test described in Section 0.0.3 was used to test for the stationarity of the data. The tests were done for the three wave periods and the results summarized in **Table 2** below.

According to the ADF test, the null hypothesis of non-stationarity is rejected whenever the p-value is less than 0.05. Hence, from the results given in **Table 2**,

the cumulative COVID-19 confirmed positive cases data for all the three waves in Kenya are second-order stationary. Therefore, the data had to be differenced twice before it was used for modeling.

3.4. Estimated Model Parameters

The negative binomial INAR (1) model was trained using the cumulative COVID-19 confirmed positive cases data for the three waves. The estimated model parameters for the three waves were independently calculated using the method of moment estimators given in Equation (33) and the results were as summarized in Table 3 below.

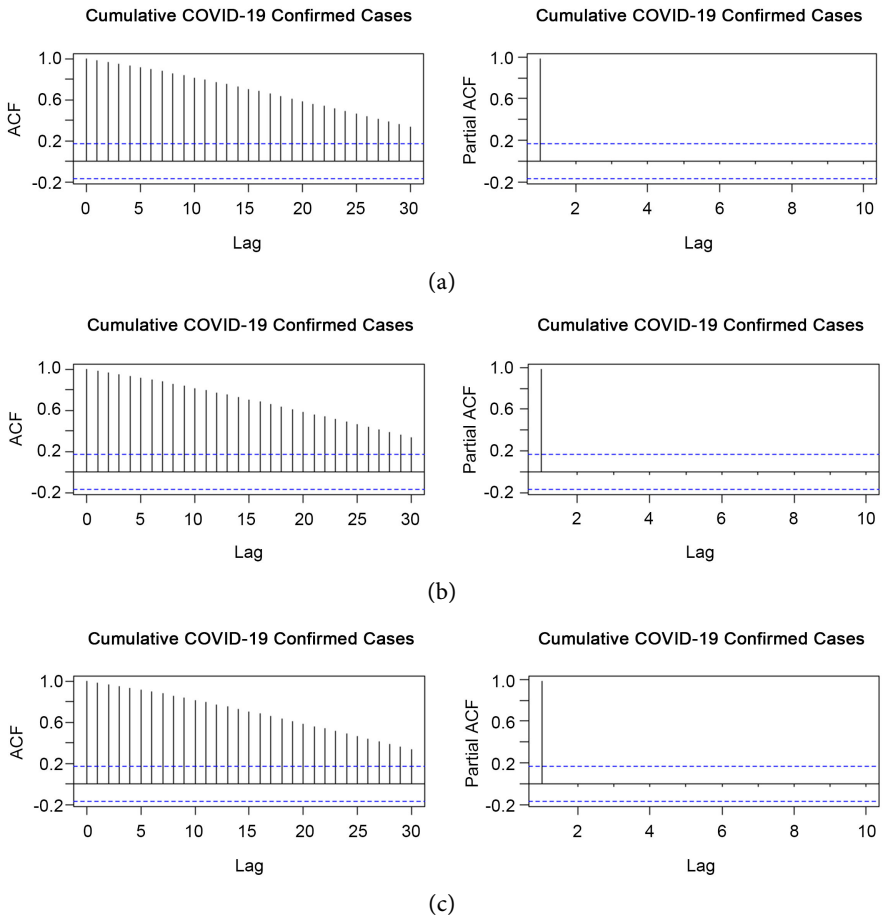


Figure 3. ACF Structures for the Kenyan Cumulative COVID-19 Train Data. (a) ACF and PACF Wave 1 plots for the period between 1/5/2020 and 11/9/2020; (b) ACF and PACF Wave 2 plots for the period between 22/9/2020 and 23/12/2020; (c) ACF and PACF Wave 3 plots for the period between 3/1/2021 and 1/3/2021.

Table 1. Descriptive statistics of the cumulative COVID-19 train data.

Wave	Period	Mean	Standard Deviation
1	5/1/2020-11/9/2020	13,346	12,587
2	22/9/2021-23/12/2020	63,049	20,422
3	3/1/2021-1/3/2021	100,958	2518

Table 2. ADF test results of the cumulative COVID-19 train data.

Period	Order of Stationarity	ADF Statistic	P-value
1/5/2020-11/9/2020	I (0)	−3.0204	0.1497
	I (1)	−0.23895	0.9900
	I (2)	−9.0810	0.0100
22/9/2020-23/12/2020	I (0)	−2.0763	0.5448
	I (1)	−1.5101	0.7782
	I (2)	−10.2450	0.0100
3/1/2021-1/3/2021	I (0)	−3.0204	0.1497
	I (1)	−0.2390	0.9900
	I (2)	−9.0810	0.0100

Table 3. Estimated negative binomial INAR (1) parameters.

Wave	Training Period	$\hat{r}_{MM}$	$\hat{p}_{MM}$	$\hat{\alpha}_{MM}$
1	1/5/2020 to 11/9/2020	2	0.9964	0.9839
2	22/9/2020 to 23/12/2020	2	0.9964	0.9772
3	3/1/2021 to 1/3/2021	5	0.9700	0.9370

3.5. Model Forecasts and Accuracy Analysis

The fitted models from the three waves were tested where 5, 10, 15 and 30-day-ahead forecasts were done and compared with their counterpart test datasets. In this section, an independent wave by wave analysis was done to ascertain the accuracy of the model for a specific wave period.

Wave 1: Forecasts

5, 10, 15 and 30-days ahead forecasts from 12th September 2020 were done on the model. The n-day-ahead forecasts for the specific periods of study were then compared with their corresponding test sample counterparts. The mean of the actual and its corresponding n-ahead forecasted values were calculated and summarized in **Table 4** below.

Figure 4 below shows the plots of the Actual and Forecasted values.

The mean absolute percentage errors (MAPEs) for the different n-day forecasts for the first wave for the 5 - 30 days ahead forecasts ranged from 0.23% to 1.78% and were as summarized in **Table 5** below.

The ACF structures of the different n-ahead forecasted values were compared with the actual sample test values. Their respective ACF and PACF plots were as shown in **Figure 5**.

The different ACF and PACF plots under **Figure 5** show that when the ACF structures of the actual sample test data and the forecasted data are compared for the different forecasts, they look very similar. However, there is need to ascertain whether the distribution functions of the different n-ahead forecasts when compared to their respective test sample counterparts are in deed statistically similar. To achieve this, the Kolmogorov-Smirnov (K-S) test was performed on the ac-

tual and forecasted datasets for the same forecast periods. The K-S test results were as summarized in Table 6 below.

From the K-S test results given in Table 6, it is clear that the 5, 10 and 15 days ahead forecasts have the same distribution function as the test sample data. However, the 30 day-ahead forecasts seem to be different from the actual test values. Hence, for the first wave, the negative-binomial INAR (1) model can adequately be used to calculate the 5, 10 and 15 day cumulative COVID-19 forecasts. However, it is not recommended to use this model for long-term modeling of the cumulative number of positive COVID-19 cases in Kenya in the first wave.

Wave 2: Forecasts

5, 10, 15 and 30-days ahead forecasts from 24th December 2020 were done on the model. The n-day-ahead forecasts for the specific periods of study were then compared with their corresponding test sample counterparts. The mean of the actual and its corresponding n-ahead forecasted values were calculated and summarized in Table 7 below.

Table 4. Wave 1 mean comparison between actual and forecasted.

				95% C.I.	
Forecast Period	n-days ahead	Actual	Forecast	L.C.I	U.C.I
12/9/2020-16/9/2020	5	36,056	35,991	33,860	38,122
12/9/2020-21/9/2020	10	36,369	36,296	34,165	38,427
12/9/2020-26/9/2020	15	36,702	36,576	34,445	38,707
12/9/2020-11/10/2020	30	37,958	37,266	35,135	39,397

Table 5. Wave 1 MAPE values for the different n-ahead forecasts.

n-days ahead	MAPE (%)
5	0.2274
10	0.2612
15	0.3809
30	1.7759

Table 6. Wave 1 kolmogorov smirnov test results.

n-days ahead	K-S Statistic	P-value
5	0.4000	0.8186
10	0.2000	0.9883
15	0.2000	0.9251
30	0.3667	0.0354

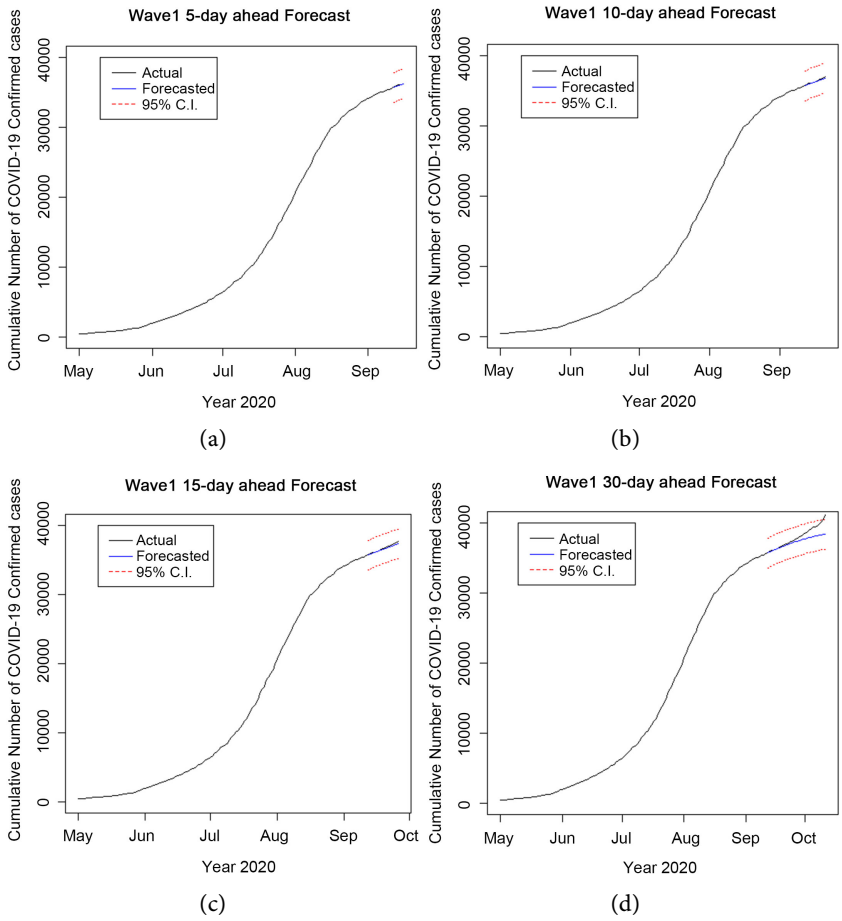
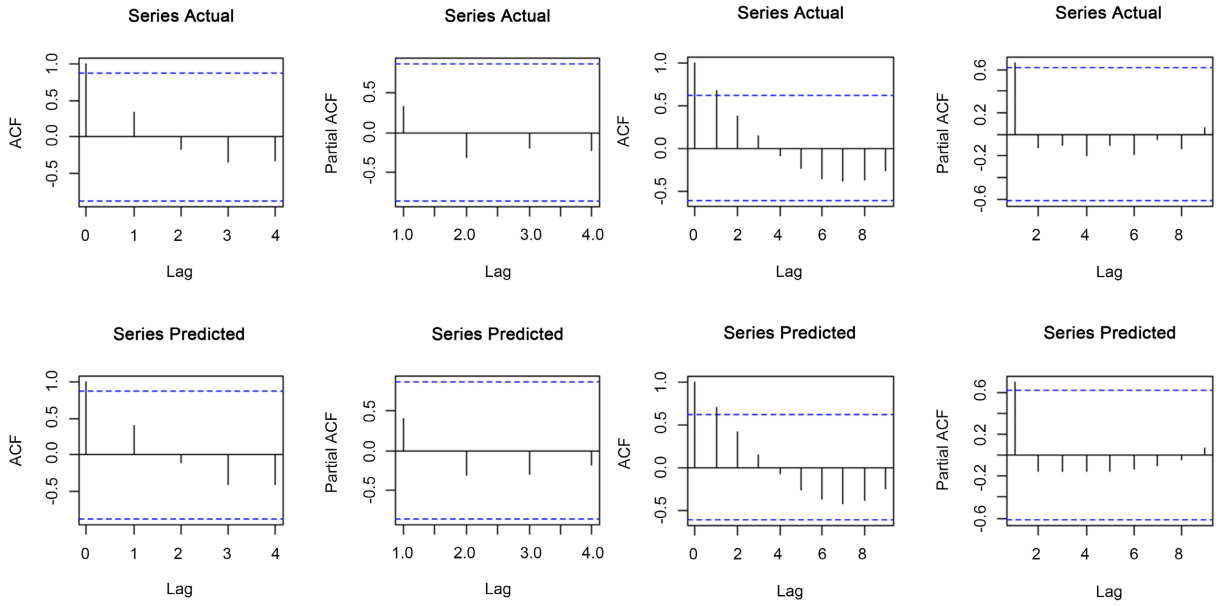


Figure 4. Wave 1 (5, 10, 15 and 30) day-ahead forecasts from 12th September 2020. (a) 5 days ahead forecast from 12/9/2020 to 16/9/2020; (b) 10 days ahead forecast from 12/9/2020 to 21/9/2020; (c) 15 days ahead forecast from 12/9/2020 to 26/9/2020; (d) 30 days ahead forecast from 12/9/2020 to 11/10/2020.



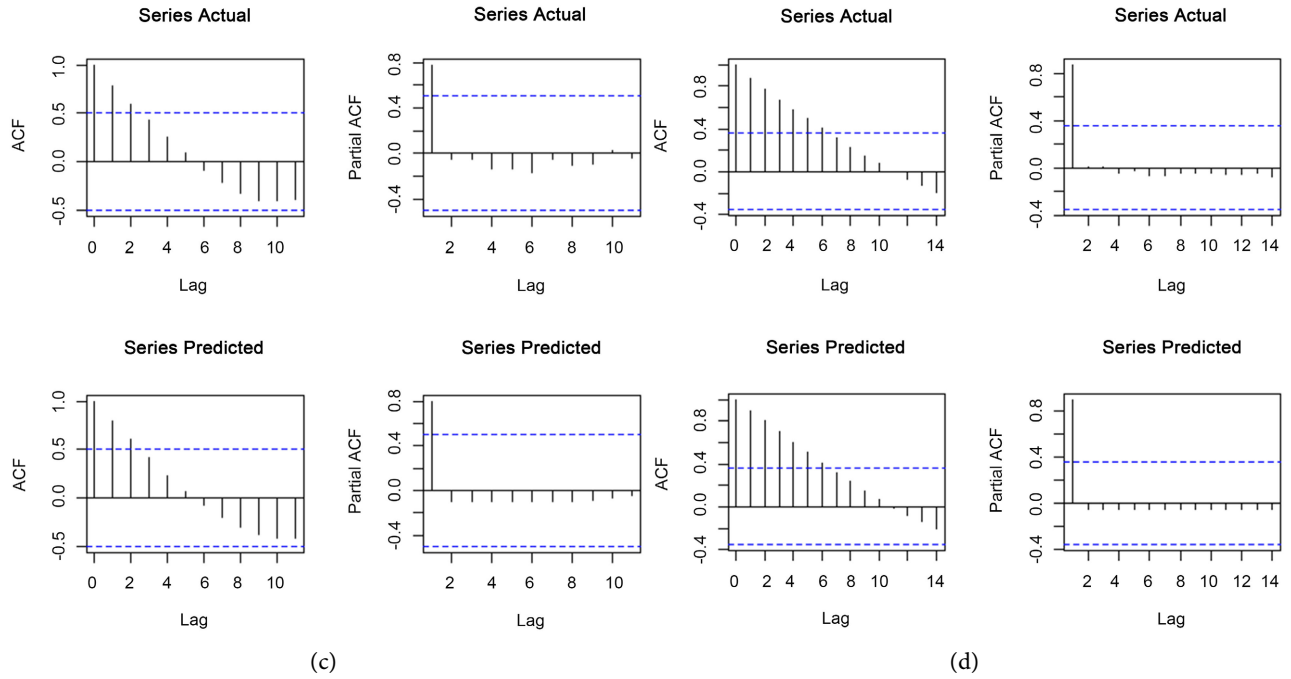


Figure 5. Wave 1 n-day-ahead Test Data ACF Structure Comparison Plots. (a) 5 days ahead ACF and PACF forecast data; (b) 10 days ahead ACF and PACF forecast data; (c) 15 days ahead ACF and PACF forecast data; (d) 30 days ahead ACF and PACF forecast data.

Table 7. Wave 2 mean comparison between actual and forecasted.

				95% C.I.	
Forecast Period	n-days ahead	Actual	Forecast	L.C.I	U.C.I
24/12/2020-28/12/2020	5	95,617	96,001	91,850	100,152
24/12/2020-2/1/2021	10	95,941	96,629	92,478	100,780
24/12/2020-7/1/2021	15	96,304	97,023	92,872	101,174
24/12/2020-22/1/2021	30	97,501	96,805	92,654	100,956

Figure 6 shows the plots of the Actual and Forecasted values.

The mean absolute percentage errors (MAPEs) for the different n-day forecasts for the second wave for the 5 - 30 days ahead forecasts ranged from 0.40% to 1.44% and was as summarized in **Table 8** below.

The ACF structures of the different n-ahead forecasted values were compared with the actual sample test values. Their respective ACF and PACF plots were as shown in **Figure 7**.

The different ACF and PACF plots under **Figure 7** show that when the ACF structures of the actual sample test data and the forecasted data are compared for the different forecasts, they look very similar. However, there is need to ascertain whether the distribution functions of the different n-ahead forecasts when compared to their respective test sample counterparts are in deed statistically similar. To achieve this, the Kolmogorov-Smirnov (K-S) test was performed on the ac-

tual and forecasted datasets for the same forecast periods. The K-S test results were as summarized in Table 9 below.

Table 8. Wave 2 MAPE values for the different n-ahead forecasts.

n-days ahead	MAPE (%)
5	0.4007
10	0.7150
15	0.7451
30	1.4400

Table 9. Wave 2 kolmogorov smirnov test results.

n-days ahead	K-S Statistic	P-value
5	0.6000	0.3571
10	0.6000	0.0525
15	0.4667	0.0755
30	0.4667	0.0025

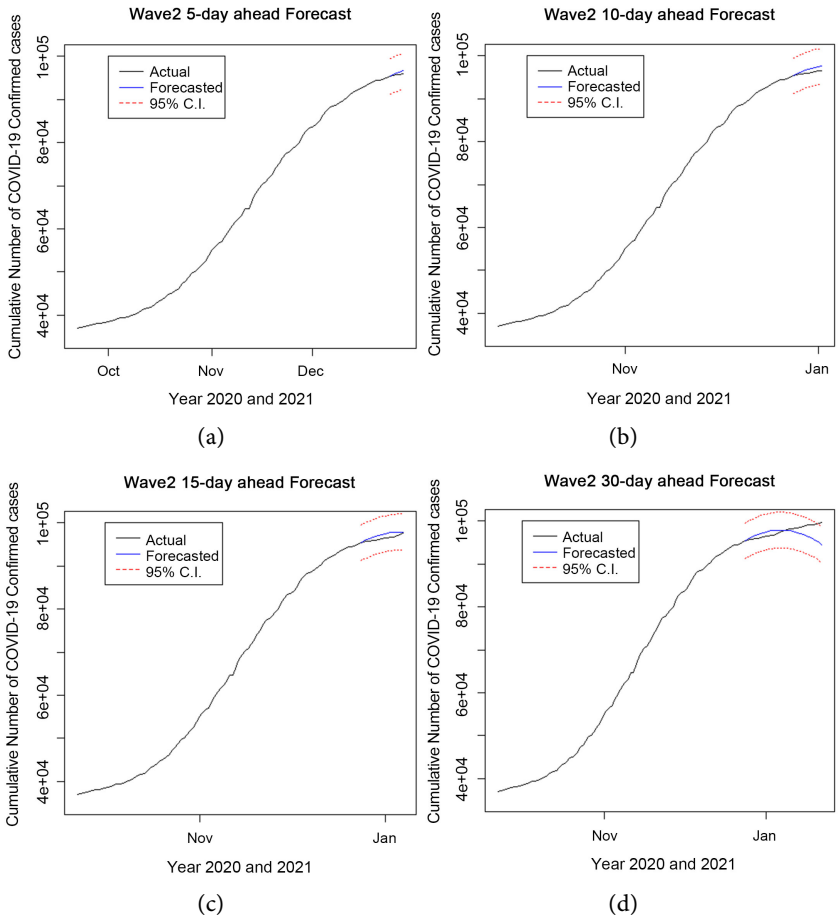


Figure 6. Wave 1 (5, 10, 15 and 30) day-ahead forecasts from 24th December 2020. (a) 5 days ahead forecast from 24/12/2020 to 28/12/2020; (b) 10 days ahead forecast from 24/12/2020 to 2/1/2021; (c) 15 days ahead forecast from 24/12/2020 to 7/1/2021; (d) 30 days ahead forecast from 24/12/2020 to 22/1/2021.

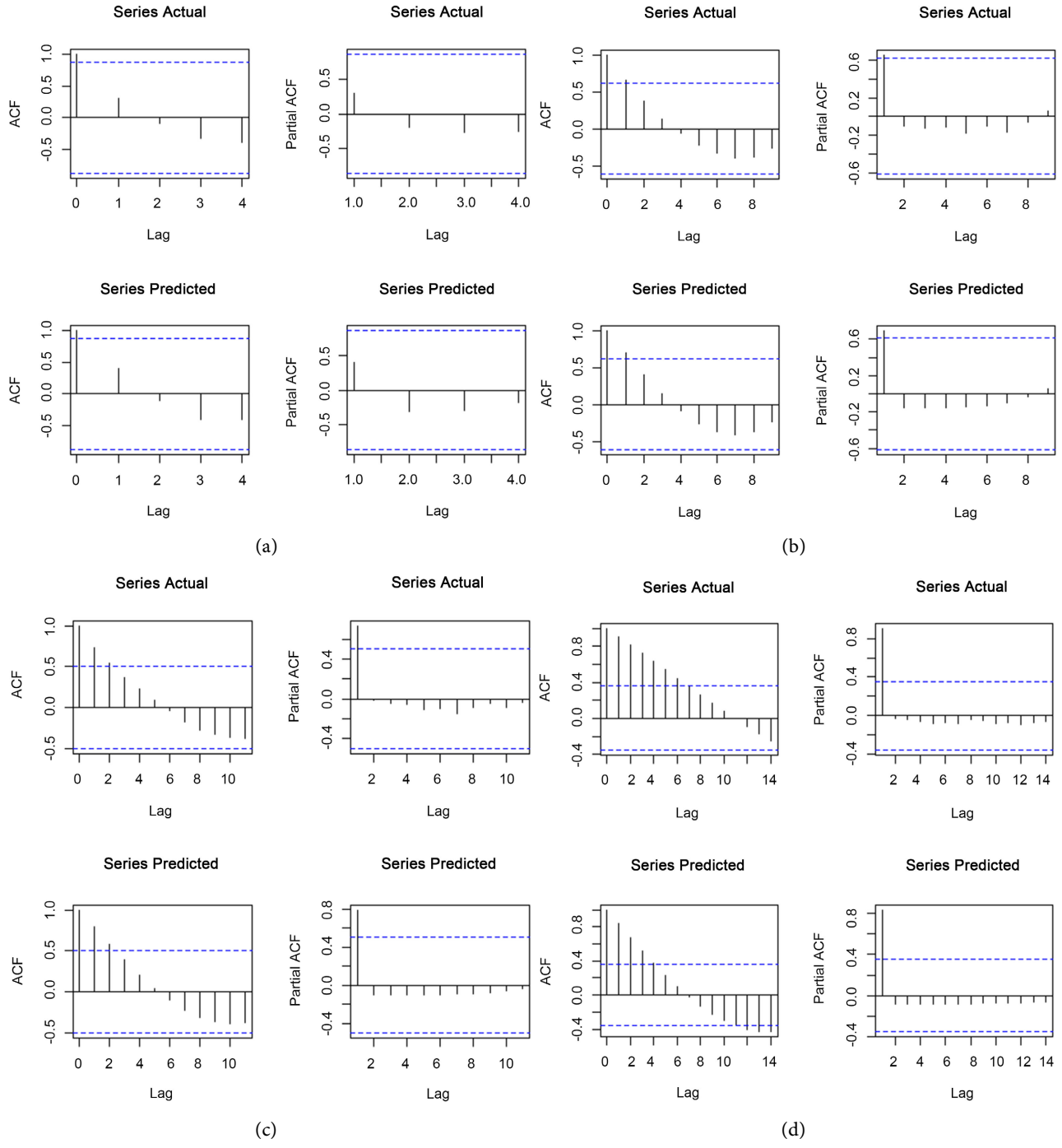


Figure 7. Wave 2 n-day-ahead Test Data ACF Structure Comparison Plots. (a) 5 days ahead ACF and PACF forecast data; (b) 10 days ahead ACF and PACF forecast data; (c) 15 days ahead ACF and PACF forecast data; (d) 30 days ahead ACF and PACF forecast data.

From the K-S test results given in Table 9, it is clear that the 5, 10 and 15 days ahead forecasts have the same distribution function as the test sample data. However, the 30 day-ahead forecasts seem to be different from the actual test values. Hence, for the second wave, the negative-binomial INAR (1) model can adequately be used to calculate the 5, 10 and 15 day cumulative COVID-19 fore-

casts. However, it is not recommended to use this model for long-term modeling of the cumulative number of positive COVID-19 cases in Kenya in the second wave.

Wave 3: Forecasts

5, 10, 15 and 30-days ahead forecasts from 2nd March 2021 were done on the model. The n-day-ahead forecasts for the specific periods of study were then compared with their corresponding test sample counterparts. The mean of the actual and its corresponding n-ahead forecasted values were calculated and summarized in Table 10 below.

Figure 8 below shows the plots of the Actual and Forecasted values.

The mean absolute percentage errors (MAPEs) for the different n-day forecasts for the first wave for the 5 - 30 days ahead forecasts ranged from 0.56% to 6.01% and were as summarized in Table 11 below.

The ACF structures of the different n-ahead forecasted values were compared with the actual sample test values. Their respective ACF and PACF plots were as shown in Figure 9.

The different ACF and PACF plots under Figure 9 show that when the ACF structures of the actual sample test data and the forecasted data are compared for the different forecasts, they look very similar. However, there is need to ascertain whether the distribution functions of the different n-ahead forecasts when compared to their respective test sample counterparts are in deed statistically similar. To achieve this, the Kolmogorov-Smirnov (K-S) test was performed on the actual and forecasted datasets for the same forecast periods. The K-S test results were as summarized in Table 12 below.

From the K-S test results given in Table 12, it is clear that the 5, 10 and 15 days ahead forecasts have the same distribution function as the test sample data. However, the 30 day-ahead forecasts seem to be different from the actual test values. Hence, for the third wave, the negative-binomial INAR (1) model can adequately be used to calculate the 5, 10 and 15 day cumulative COVID-19 forecasts. However, it is not recommended to use this model for long-term modeling of the cumulative number of positive COVID-19 cases in Kenya in the third wave.

Table 10. Wave 3 mean comparison between actual and forecasted.

				95% C.I.	
Forecast Period	n-days ahead	Actual	Forecast	L.C.I	U.C.I
2/3/2021-6/3/2021	5	107,175	106,572	105,824	107,220
2/3/2021-11/3/2021	10	108,424	107,125	106,477	107,773
2/3/2021-16/3/2021	15	110,026	107,744	107,096	108,392
2/3/2021-31/3/2021	30	117,446	110,001	109,353	110,649

Table 11. Wave 3 MAPE values for the different n-ahead forecasts.

n-days ahead	MAPE (%)
5	0.5606
10	1.1881
15	2.0414
30	6.0088

Table 12. Wave 3 kolmogorov smirnov test results.

n-days ahead	K-S Statistic	P-value
5	0.6000	0.3571
10	0.6000	0.0525
15	0.4667	0.0755
30	0.5000	0.0009

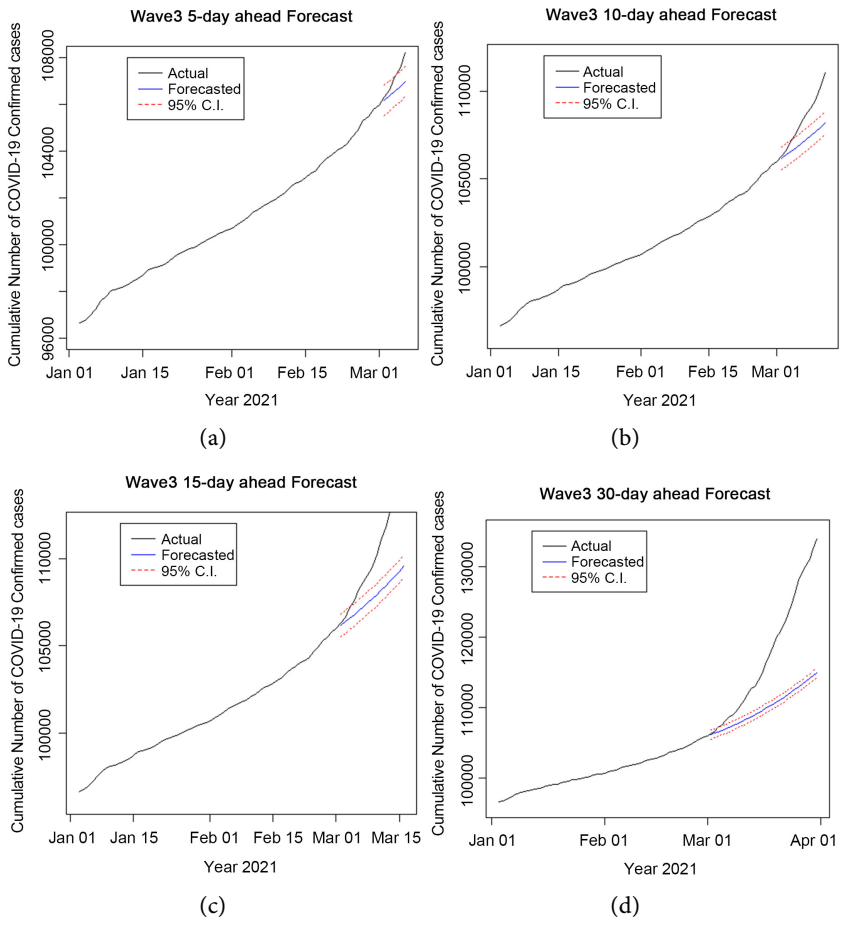


Figure 8. Wave 3 (5, 10, 15 and 30) day-ahead forecasts from 4th March 2021. (a) 5 days ahead forecast from 2/3/2021 to 6/3/2021; (b) 10 days ahead forecast from 2/3/2021 to 11/3/2021; (c) 15 days ahead forecast from 2/1/2021 to 16/3/2021; (d) 30 days ahead forecast from 2/1/2021 to 31/1/2021.

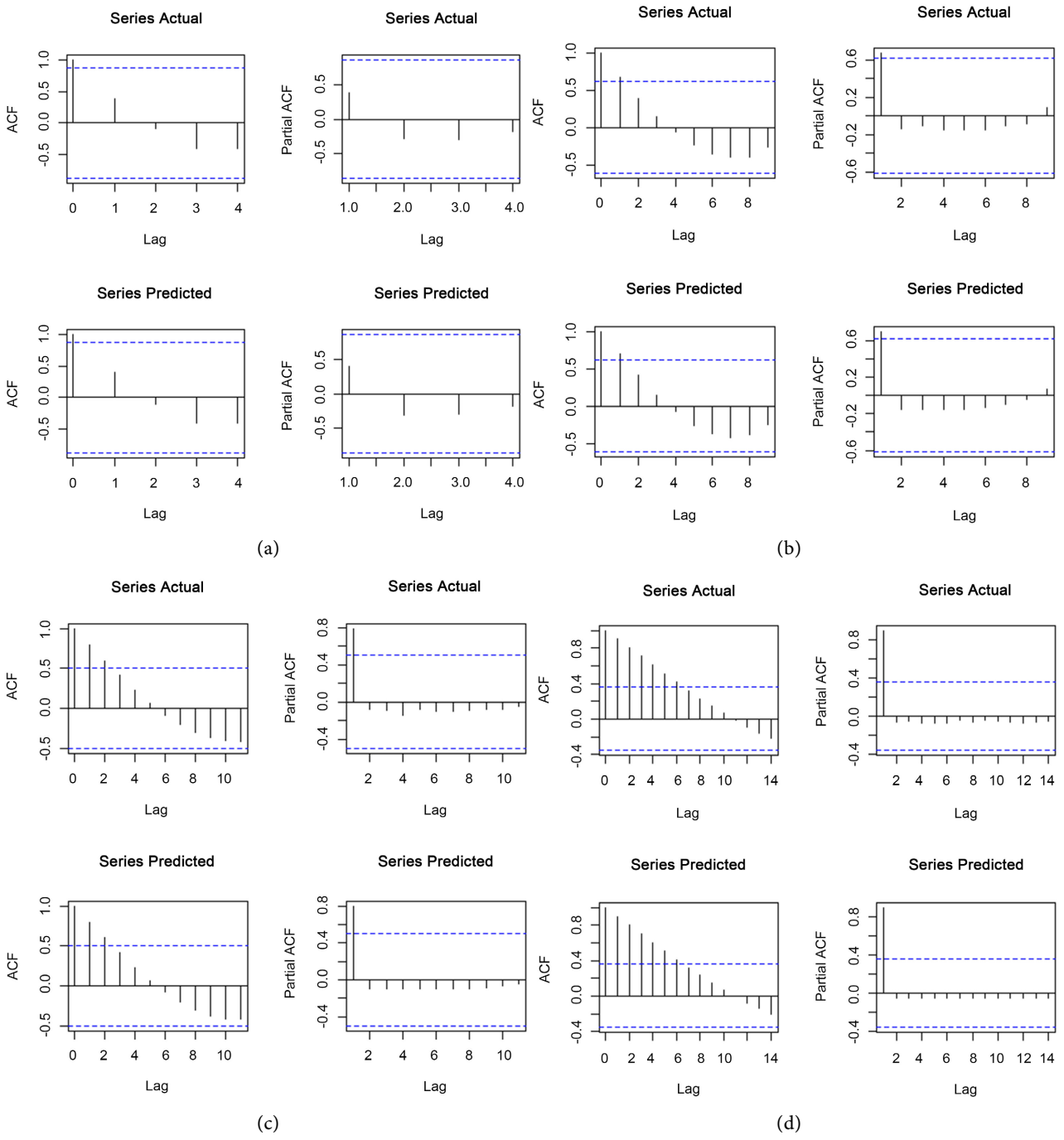


Figure 9. Wave 3 n-day-ahead Test Data ACF Structure Comparison Plots. (a) 5 days ahead ACF and PACF forecast data; (b) 10 days ahead ACF and PACF forecast data; (c) 15 days ahead ACF and PACF forecast data; (d) 30 days ahead ACF and PACF forecast data.

4. Conclusions

In this paper, our main interest was in modeling the cumulative number of positive COVID-19 cases in Kenya using an appropriate INAR (1) model. In our analysis, the data was sub-divided to cater for the three COVID-19 waves Kenya has suffered so far. The cumulative positive COVID-19 cases data used was

found to be over-dispersed and hence the negative-binomial INAR (1) model was considered to be a more appropriate consideration for modeling the integer-valued dataset. In addition to this, the cumulative COVID-19 data considered to train the model was found to be second order stationary.

5, 10, 15 and 30 day-ahead out of sample forecasts were performed on the data for the three waves independently. To ascertain the model accuracy of the forecasted values, different parameters were considered. First and foremost, the ACF structures for the out of sample forecasts were analyzed and compared with those of the test dataset. The ACF and PACF plots of the n-days ahead forecasted values were found to be identical to their counterpart test values for all three waves. Hence, the INAR (1) model seemed to be appropriate to model the different COVID-19 waves regardless of the number of days ahead considered.

To be certain of the similarity between the forecasted values and the test values, the distribution functions of the different n-ahead forecasts were compared to those of the test dataset. This was achieved using the two-sample Kolmogorov Smirnov test. The results of this test showed that the 5, 10, and 15 day-ahead COVID-19 forecasts for the three different waves were statistically similar to each other at 5% level of significance. However, the 30 day ahead forecasts were found to be statistically different from each other. It was therefore concluded that the negative binomial INAR (1) model is appropriate for short-term COVID-19 forecasting.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Al-Osh, M.A. and Alzaid, A.A. (1987) First-Order Integer-Valued Autoregressive (INAR (1)) Process. *Journal of Time Series Analysis*, **8**, 261-275. https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

[2] Al-Qaness, M.A.A., Ewees, A.A., Fan, H. and Abd El Aziz, M. (2020) First-Order Integer-Valued Autoregressive (INAR (1)) Process. *Journal of Clinical Medicine*, **9**, 674. https://doi.org/10.3390/jcm9030674

[3] Argawu, A.S. (2020) Modeling and Forecasting of COVID-19 New Cases in the Top 10 Infected African Countries Using Regression and Time Series Models. Cold Spring Harbor Laboratory Press, medRxiv. https://doi.org/10.1101/2020.09.23.20200113

[4] Ceylan, Z. (2020) Estimation of COVID-19 Prevalence in Italy, Spain, and France. *Science of the Total Environment*, **729**, Article ID: 138817. https://doi.org/10.1016/j.scitotenv.2020.138817

[5] Massey Jr, F.J. (2020) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, **46**, 68-78. https://doi.org/10.1080/01621459.1951.10500769

[6] McKenzie, E. (1984) Some Simple Models for Discrete Variate Time Series. *Water Resources & Bulletin*, **21**, 645-650.

https://doi.org/10.1111/j.1752-1688.1985.tb05379.x

[7] Sahai, A.K., Rath, N., Sood, V. and Singh, M.P. (2020) ARIMA Modelling & Forecasting of COVID-19 in Top Five Affected Countries. *Diabetes & Metabolic Syndrome*: *Clinical Research & Reviews*, **14**, 1419-1427. https://doi.org/10.1016/j.dsx.2020.07.042

[8] Şahin, U. and Şahin, T. (2020) Forecasting the Cumulative Number of Confirmed Cases of COVID-19 in Italy, UK and USA Using Fractional Nonlinear Grey Bernoulli Model. *Chaos*, *Solitons & Fractals*, **138**, Article ID: 109948. https://doi.org/10.1016/j.chaos.2020.109948

[9] Takele, R. (2020) Stochastic Modelling for Predicting COVID-19 Prevalence in East Africa Countries. *Infectious Disease Modelling*, **5**, 598-607. https://doi.org/10.1016/j.idm.2020.08.005

[10] Triacca, M. and Triacca, U. (2021) Forecasting the Number of Confirmed New Cases of COVID-19 in Italy for the Period from 19 May to 2 June 2020. *Infectious Disease Modelling*, **6**, 362-369. https://doi.org/10.1016/j.idm.2021.01.003

[11] Zeroual, A., Harrou, F., Dairi, A. and Sun, Y. (2020) Deep Learning Methods for Forecasting COVID-19 Time-Series Data: A Comparative Study. *Chaos*, *Solitons & Fractals*, **140**, Article ID: 110121. https://doi.org/10.1016/j.chaos.2020.110121